POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

An overview on the Italian sustainability reporting: the Social pillar

Tesi di Laurea Magistrale in
Mathematical Engineering - Ingegneria Matematica

Author: **Stefano Mezzera**

Student ID: 953578
Advisor: Prof. Francesca Grassetti
Academic Year: 2021-22

Abstract

Legislative Decree No. 254/2016, introduced the obligation to publish a Non-Financial Statement (DNF) for Italian listed companies.
In order to make sustainability reporting homogeneous, the GRI aims to define standards for DNF reporting. In this paper we will discuss the data published by 164 FTSE All share companies, covering the year 2020, related to the Social pillar.
We will highlight the many discrepancies that occur in the reporting of the various companies and present the first exploratory analyses of the GRI indices of the 400 series (social issues), relating to recruitment (401), occupational health and safety (403), staff training (404), and diversity and equal opportunities (405).
Finally, through properties of complex networks, possible relationships between these companies will be analysed.

Keywords: ESG, DNF, Complex networks, GRI standards, Social pillar, Similarity measures.

Abstract in lingua italiana

Il decreto legislativo n.254/2016, ha introdotto l'obbligo di pubblicare una Dichiarazione di carattere Non Finanziario (DNF) per le società italiane quotate. Per rendere omogenea la reportistica in ambito di sostenibilità, la GRI ha l'obiettivo di definire standard di rendicontazione delle DNF.
In questo lavoro tratteremo i dati pubblicati da 164 società del FTSE All share, riguardanti l'anno 2020, relative al pillar Social.
Verranno sottolineate le numerose difformità che si presentano nelle rendicontazioni delle varie società e si presenteranno le prime analisi esplorative degli indici GRI della serie 400 (temi sociali), relativi alle assunzioni (401), alla salute e alla sicurezza sul luogo di lavoro (403), alla formazione del personale (404), e a diversità e pari opportunità (405).
Infine, tramite proprietà delle reti complesse verranno analizzati possibili relazioni tra queste società.

Parole chiave: ESG, DNF, Reti complesse, standard GRI , pillar Sociale , Misure di similarità.

Contents

Acknowledgements **iii**

Abstract **i**

Abstract in lingua italiana **iii**

Contents **v**

1 Similarity measures **1**
1.1 Concept of distance and similarity 1
1.1.1 Common distances and similarities 2
1.2 Norm of difference of adjacency matrices 5
1.2.1 Main norms 6
1.3 Hierarchical clustering 8
1.3.1 Hierarchical clustering techniques 8
1.3.2 Algorithm: Hierarchical clustering 8
1.3.3 The interpretation of a dendogram 10
1.3.4 The choice of the number of clusters 12

2 The structure of complex networks **15**
2.1 Definitions and notations 15
2.1.1 Node degree 16
2.1.2 Correlated networks 16
2.1.3 Distance and diameter 18
2.1.4 Clustering coefficient 18
2.1.5 Weighted networks 19
2.2 Network models 22
2.2.1 Random networks: Erdös - Rényi 22
2.2.2 Scale-free networks: Barabási – Albert 23

2.2.3 Small-world networks: Watts - Strogatz . . . 24
2.3 Node centralities . . . 26
2.4 Community detection . . . 26
2.4.1 The method of Modularity optimization . . . 27
2.4.2 Finding communities by means of random walkers . . . 30
2.5 Link prediction . . . 32
2.5.1 Similarity-based algorithms . . . 33

3 Database report: Social pillar **35**
3.1 ESG and Social pillar . . . 35
3.2 Database costruction . . . 35
3.2.1 GRI standard indices . . . 40
3.3 Database analysis . . . 42

4 Network analysis **65**
4.1 Data Prepocessing . . . 65
4.2 Matrix of weights . . . 67
4.3 Network analysis . . . 70
4.4 Candidate communities . . . 71
4.5 Results . . . 75

Bibliography **79**

List of Figures **83**

List of Tables **85**

1 | Similarity measures

1.1. Concept of distance and similarity

The assessment of how different two objects are, is a key requirement in many artificial intelligence and machine learning methods. Distance and similarity are mathematical tools very useful in these kind of algorithms.
Distance are functions which assign a numerical value to each pair of objects in a given domain.
This value can be interpreted as a measure of how much an object and another are whether alike or not: two very similar objects will be assigned a low distance, instead a larger value to two dissimilar ones.
A complementary concept to the distance one, is the similarity, maps which assign higher value to similar objects and lower to different couples.

Definition 1.1. ***Distance***
Let $\Omega \neq \emptyset$. A function $d : \Omega \times \Omega \to [0, +\infty)$ is said to be a distance if:

1. $d(x, y) \geq 0 \quad \forall x, y \in \Omega$*(Non-negativity).*
2. $d(x, y) = 0 \iff x = y$ *(Identity).*
3. $d(x, y) = d(y, x)$ *(Simmetry).*
4. $d(x, z) \leq d(x, y) + d(y, z) \quad \forall x, y, z \in \Omega$ *(Triangle inequality).*

Definition 1.2. ***Similarity***
Let $\Omega \neq \emptyset$. A function $s : \Omega \times \Omega \to [0, u]$, where u is an upper bound, i.e, the maximum similarity value, is said to be a similarity if the following properties are satisfied:

1. $s(x, y) = u \iff x = y$ *(Identity).*
2. $s(x, y) = s(y, x)$ *(Simmetry).*

Proposition 1.1. ***Maximum similarity value***

The usual value assigned to the maximum similarity value is $u = 1$.

Considering the previous definitions of distance and similarity, it is possible to think of these two metrics as a complementary concept, moreover we can build a similarity associated to any distance[27]:

Proposition 1.2. ***Similarity associated to distance***
For each distance d, we can define its associated similarity as follow:

$$s_d(x, y) = \frac{u}{1 + d(x, y)}$$

It is important to stress the fact that distance can be undounded, whereas similarity range in the compact set $[0, u]$.
Moreover there is not an equivalent property to the triangle inequality for similarity.

1.1.1. Common distances and similarities

Consider two data points, $X = [x_1, x_2, ..., x_p]$ and $Y = [y_1, y_2, ..., y_p]$. We now define some common distances.

Definition 1.3. ***Euclidean distance***

$$d(X, Y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

Definition 1.4. ***Minkowski distance***

$$d(X, Y) = \left(\sum_{i=1}^{p} (x_i - y_i)^r \right)^{1/r}$$

where r is a fixed parameter.
Some important cases are:

1. $r = 1$*: Manhattan distance.*
2. $r = 2$*: Euclidean distance.*
3. $r = \infty$*: Chebychev distance.*

Definition 1.5. ***Manhattan distance***

$$d(X,Y) = \sum_{i=1}^{p} |x_i - y_i|$$

Definition 1.6. *Chebychev distance*

$$d(X,Y) = \max_i |x_i - y_i|$$

To better understand the relations between Minkowski distances, in Figure 1.2[19] it is possible to see what they actually measure in a two-dimensional space.

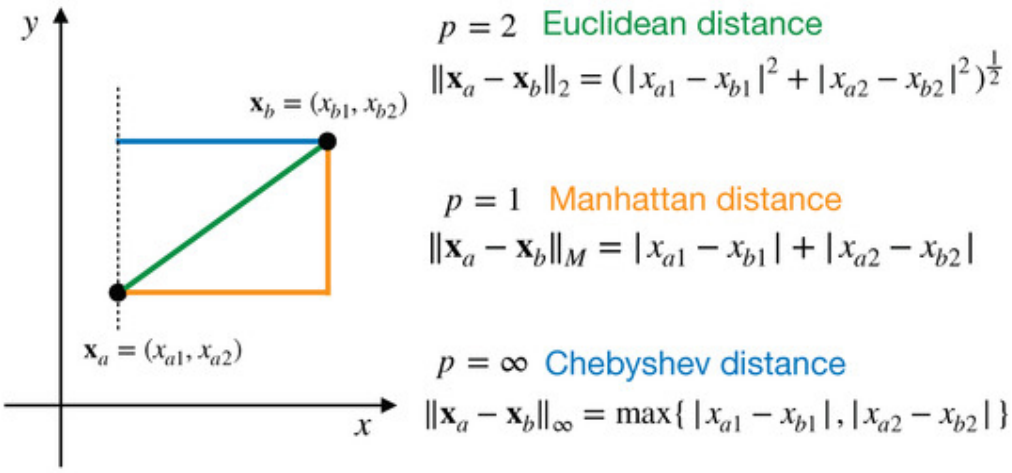


Figure 1.1: Minkowski distances in a two dimensional space

Definition 1.7. *Canberra distance*

$$d(X,Y) = \sum_{i=1}^{p} \frac{|x_i - y_i|}{|x_i| + |y_i|}$$

Definition 1.8. *Mahalanobis distance*

$$d(X,Y) = \sqrt{(X-Y)\,\sigma(X,Y)^{-1}\,(X-Y)^T}$$

where $\sigma(X,Y)$ *is the covariance matrix.*

Definition 1.9. *Jaccard similarities*

- ***Jaccard similarity for finite sets:***

$$d(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

where the operator /./ stand for the cardinality of the set.

- ***Jaccard similarity for binary vectors:***
Given two binary vectors $X = [x_1, x_2, ..., x_p]$ *and* $Y = [y_1, y_2, ..., y_p]$*, i.e, the entrance* $x_i, y_i \in \{0, 1\}$

$$d(X,Y) = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$$

where:

  - M_{10} *is the number of entrance where X is 1 and Y is 0.*
  - M_{01} *is the number of entrance where X is 0 and Y is 1.*
  - M_{11} *is the number of entrance where X is 1 and Y is 1.*

The different similarity measures presented up to now focus on the magnitude of the observations.
By the way, can be useful to introduce a metric, which takes into consideration the shape of the profile. The key role in this framework is played by the correlation, in fact, it is possible to introduce a correlation-based distance[20].

Definition 1.10. ***Correlation***

$$r(X,Y) = \frac{\sum_{i=1}^{p}(x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{p}(x_i - \overline{X})^2 \sum_{i=1}^{p}(y_i - \overline{Y})^2}}$$

$r(X,Y) \in [-1, 1]$.

Definition 1.11. ***Pearson correlation distance***

$$d(X,Y) = 1 - r(X,Y)$$

Notice that, the Pearson correlation distance ranges in $[0, 2]$.
We have $d(X,Y) = 0$ if X and Y are a linear transformation with negative slope, and $d(X,Y) = 2$ if the slope is positive.
In case X and Y are uncorrelated then $d(X,Y) = 1$.

1.2. Norm of difference of adjacency matrices

For the general purpose of this study, in order to compare two different graphs, as a first step, we have to introduce a method to measure the distance between them.
Since the graph topology is completely identified by its adjacency matrix, we now consider some matrix norms to analyze the network distances [30].
Given two networks $\mathcal{G}_1 = (\mathcal{N}_1, \mathcal{L}_1)$ and $\mathcal{G}_2 = (\mathcal{N}_2, \mathcal{L}_2)$, consider the corresponding adjacency matrices $A_1 = (a^1_{ij})$ and $A_2 = (a^2_{ij})$.
If the networks $\mathcal{G}_1$, $\mathcal{G}_2$ don't have the same set of nodes, $\mathcal{N}_1 \neq \mathcal{N}_2$, it is enough to consider the union of these sets, $\mathcal{N} = \mathcal{N}_1 \cup \mathcal{N}_2$.
In this scenario it is necessary to change the adjacency matrices:

- A_k dimension becomes $N \times N$, where $N = N_1 + N_2 - |\mathcal{N}_1 \cap \mathcal{N}_2| \quad k = 1, 2$.
- if node $j \in \mathcal{N} \setminus \mathcal{N}_k$ add to A_k the j-th row and the j-th column setting to 0 all values.

$$A_1 = \begin{bmatrix} a^1_{11} & \cdots & a_{1N^1_1} \\ \vdots & \ddots & \vdots \\ \vdots & & \vdots \\ a^1_{N_1 1} & \cdots & a^1_{N_1 N_1} \end{bmatrix} \qquad A_2 = \begin{bmatrix} a^2_{11} & \cdots & a^2_{1N_2} \\ \vdots & \ddots & \vdots \\ a^2_{N_2 1} & \cdots & a^2_{N_2 N_2} \end{bmatrix}$$

$$A_1 = \underbrace{\begin{bmatrix} a^1_{11} & \cdots & a^1_{N_1 1} & 0 \cdots & 0 \\ \vdots & \ddots & \vdots & 0 \cdots & 0 \\ \vdots & & \vdots & 0 \cdots & 0 \\ a^1_{N_1 1} & \cdots & a^1_{N_1 N_1} & 0 \cdots & 0 \\ 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & \cdots & 0 \end{bmatrix}}_{N} \begin{matrix} \left.\vphantom{\begin{matrix} a \\ a \\ a \\ a \end{matrix}}\right\} N_1 \\ \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} N - N_1 \end{matrix}$$

$$A_2 = \underbrace{\begin{bmatrix} 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & a^2_{11} & \cdots & a^2_{N_2 1} \\ 0 & \cdots & \vdots & & \vdots \\ 0 & \cdots & a^2_{N_2 1} & \cdots & a^2_{N_2 N_2} \end{bmatrix}}_{N} \begin{matrix} \left.\vphantom{\begin{matrix} a \\ a \\ a \\ a \end{matrix}}\right\} N - N_2 \\ \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} N_2 \end{matrix}$$

1.2.1. Main norms

Definition 1.12. ***Euclidean distance***

$$d(\mathcal{G}_1, \mathcal{G}_2) = \sqrt{\sum_{i,j \in \mathcal{N}} \left(a_{ij}^1 - a_{ij}^2\right)^2}$$

Definition 1.13. ***Manhattan distance***

$$d(\mathcal{G}_1, \mathcal{G}_2) = \sum_{i,j \in \mathcal{N}} |a_{ij}^1 - a_{ij}^2|$$

Definition 1.14. ***Canberra distance***

$$d(\mathcal{G}_1, \mathcal{G}_2) = \sum_{i,j \in \mathcal{N}} \frac{|a_{ij}^1 - a_{ij}^2|}{|a_{ij}^1| + |a_{ij}^2|}$$

If $a_{ij}^1 = a_{ij}^2 = 0$, then we set $|a_{ij}^1| + |a_{ij}^2| = 1$.

Definition 1.15. ***Jaccard distance***

$$d(\mathcal{G}_1, \mathcal{G}_2) = 1 - J(A_1, A_2)$$

where $J(A_1, A_2)$ is the common Jaccard similarity:

$$J(A_1, A_2) = \frac{|A_1 \cap A_2|}{|A_1 \cap A_2|} = \frac{M_{11}}{M_{10} + M_{01} + M_{11}}$$

where:

- *M_{10} is the number of entrance where A_1 is 1 and A_2 is 0.*
- *M_{01} is the number of entrance where A_1 is 0 and A_2 is 1.*
- *M_{11} is the number of entrance where A_1 is 1 and A_2 is 1.*

The Jaccard distance can be generalized in order to handle weighted networks:

Definition 1.16. ***Weighted Jaccard distance***

$$d(\mathcal{G}_1, \mathcal{G}_2) = 1 - J_W(A_1, A_2)$$

where $J_W(A_1, A_2)$ *is defined as follow:*

$$J_W(A_1, A_2) = \begin{cases} \frac{\sum_{i,j \in \mathcal{N}} \min(a_{ij}^1, a_{ij}^2)}{\sum_{i,j \in \mathcal{N}} \max(a_{ij}^1, a_{ij}^2)}, & \textit{if } \sum_{i,j \in \mathcal{N}} \max(a_{ij}^1, a_{ij}^2) > 0 \\ 1, & \textit{if } \sum_{i,j \in \mathcal{N}} \max(a_{ij}^1, a_{ij}^2) < 0 \end{cases}$$

1.3. Hierarchical clustering

1.3.1. Hierarchical clustering techniques

Clustering refers to a set of techniques for finding sub groups in a data set.
The objective is to split the data set into distinct groups, such that observation within the group are similar, vice versa, data in different clusters are different.
More precisely, the aim is to maximize the inter cluster distance, i.e, the distance between different groups, and minimizing the intra cluster distance, i.e, the distance among observations in the same cluster.

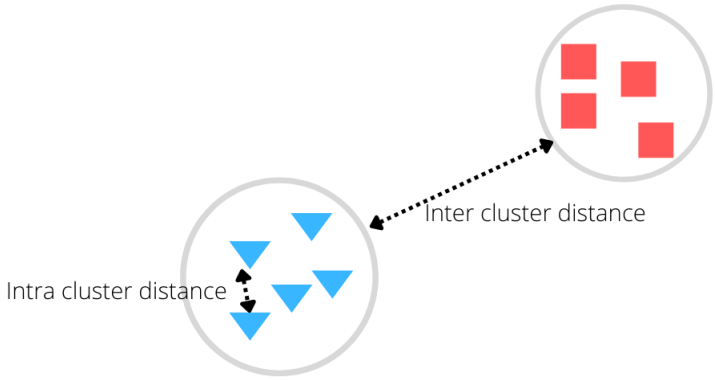


Figure 1.2: Intra and inter cluster distances

To quantify this similarity, we will use the already discussed distances and similarity measures.
For our purpose we focus on one of the best known approaches: the hierarchical clustering. One of the main features of these algorithms is the fact that it is not required the knowledge of the number of clusters in advance, in fact, thanks to a visual representation, the dendogram, it is possible to visualize at once all possible number K of clusters we can obtain, with the trivial cases, $K = 1$ the whole data set, and $K = n$ each observation is a cluster.

1.3.2. Algorithm: Hierarchical clustering

In this section we present the hierarchical clustering algorimtm, which will produce the dendogram. The output will be uniquely determined by an a priori choice, the distance and the linkage. [16][21]
The pseudo-algorithms is shown in Algorithm 1.1.

Algorithm 1.1 Hierarchical clustering

1: Begin with n groups,i.e, each observation is classified as a cluster.
2: Compute a distance matrix of all $\binom{n}{2} = n(n-1)/2$ pairwise similarities.
3: **for** $i = n, n-1, ..., 2$ **do**
4: Evaluate all pairwise inter-cluster similarity among the i groups.
5: Identify the most similar couple and merge the two.
6: The distance between the two merged cluster is the height in the dendogram at which the connection appears.
7: Update the pairwise distance for the new $i-1$ clusters.
8: **end for**

At each iteration it is necessary to update the distance matrix, considering that the number of clusters is decreasing.
We have, thus, to extend the notion of distance between clusters, since clusters can contain multiple observations. This is achieved by the notion of linkage.
We consider now the four most common typologies of linkages:
Let x be an observation in the data set, and consider the cluster, obtained via the execution of the algorithm, by merging two existing clusters A and B.

Definition 1.17. ***Single Linkage***
The Single linkage algorithm outputs clusters formed by merging clusters with minimal inter cluster distance, and the update of the distance matrix, at each step, is computed as follow:

$$d(x, A \cup B) = \min\{d(x, A), d(x, B)\}$$

Definition 1.18. ***Complete Linkage***
The Complete linkage algorithm outputs clusters formed by merging clusters with maximal inter cluster distance, and the update of the distance matrix, at each step, is computed as follow:

$$d(x, A \cup B) = \max\{d(x, A), d(x, B)\}$$

Definition 1.19. ***Average Linkage***
The Average linkage treats the distance between two clusters as the average between all pairs of observations belonging to each group, and the update of the distance matrix, at each step, is computed as follow:

$$d(x, A \cup B) = \frac{d(x, A) + d(x, B)}{|A \cup B|}$$

Definition 1.20. ***Error Sum of Squares(ESS)***
The Error Sum of Squares(ESS), is the sum of the squared differences between each observation and its group's mean:

$$ESS = \sum_{i=1}^{N} x_i^2 - \frac{1}{N}(\sum_{i=1}^{N} x_i)^2$$

Definition 1.21. ***Centroid Linkage***
The Centroid linkage algorithm is based on minimizing the increase, in terms of ESS, caused by the merge of two clusters, and the update of the distance matrix, at each step, is computed as follow:

$$d(x, A \cup B) = \frac{d(x, A) + d(x, B)}{2}$$

This last linkage algorithm, presents a major drawback, an inversion can occur,i.e, two clusters are merged at a height below either of the individual clusters in the dendrogram. This can lead to some issues both in visualization and in the interpretation of the dendrogram.

1.3.3. The interpretation of a dendogram

Once obtained the dendogram, the most important and somehow difficult choice is the selection of the number of clusters. We will take into consideration a simple example to highlight the procedure.
Consider the following toy data set, represented in Figure 1.3, consisting of 15 observations in a two dimensional space, divided in three different classes, which are underlined by different colors.

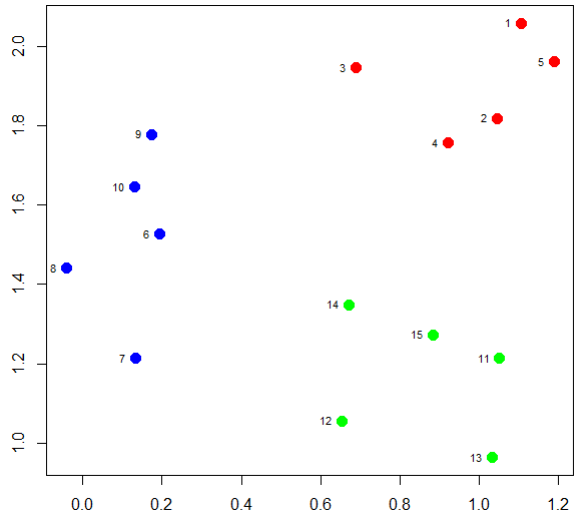


Figure 1.3: dataset

If we execute the hierarchical clustering algorithm on our data, choosing the Euclidean distance with the complete linkage, Figure 1.4 is what we obtain.
Before entering in the details of the methodology to select a proper number of cluster, it is interesting exploiting the algorithm steps.In Figure 1.5, the first four steps are visually proposed: the algorithm starts with $K = 15$ clusters, at first step the most similar, observation 1 and 5 are merged, and the update of the distance matrix is computed. In the dendogram appear the connection of the two leaf at the height corresponding to the distance between them. This procedure is iterated until $K = 1$.

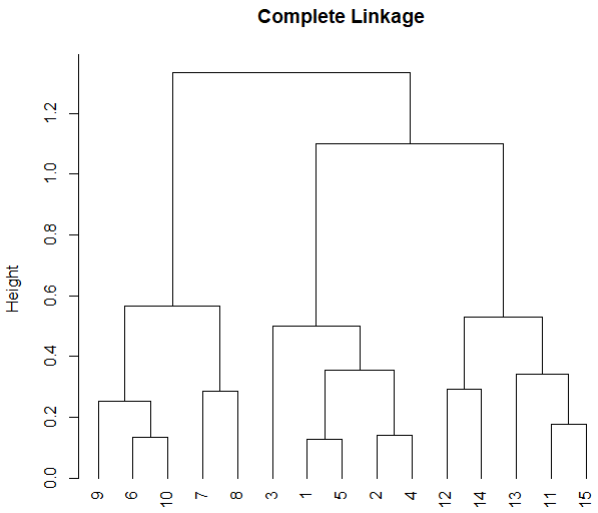


Figure 1.4: Dendogram, output of the hierarchical clustering algorithm on the toy data set.

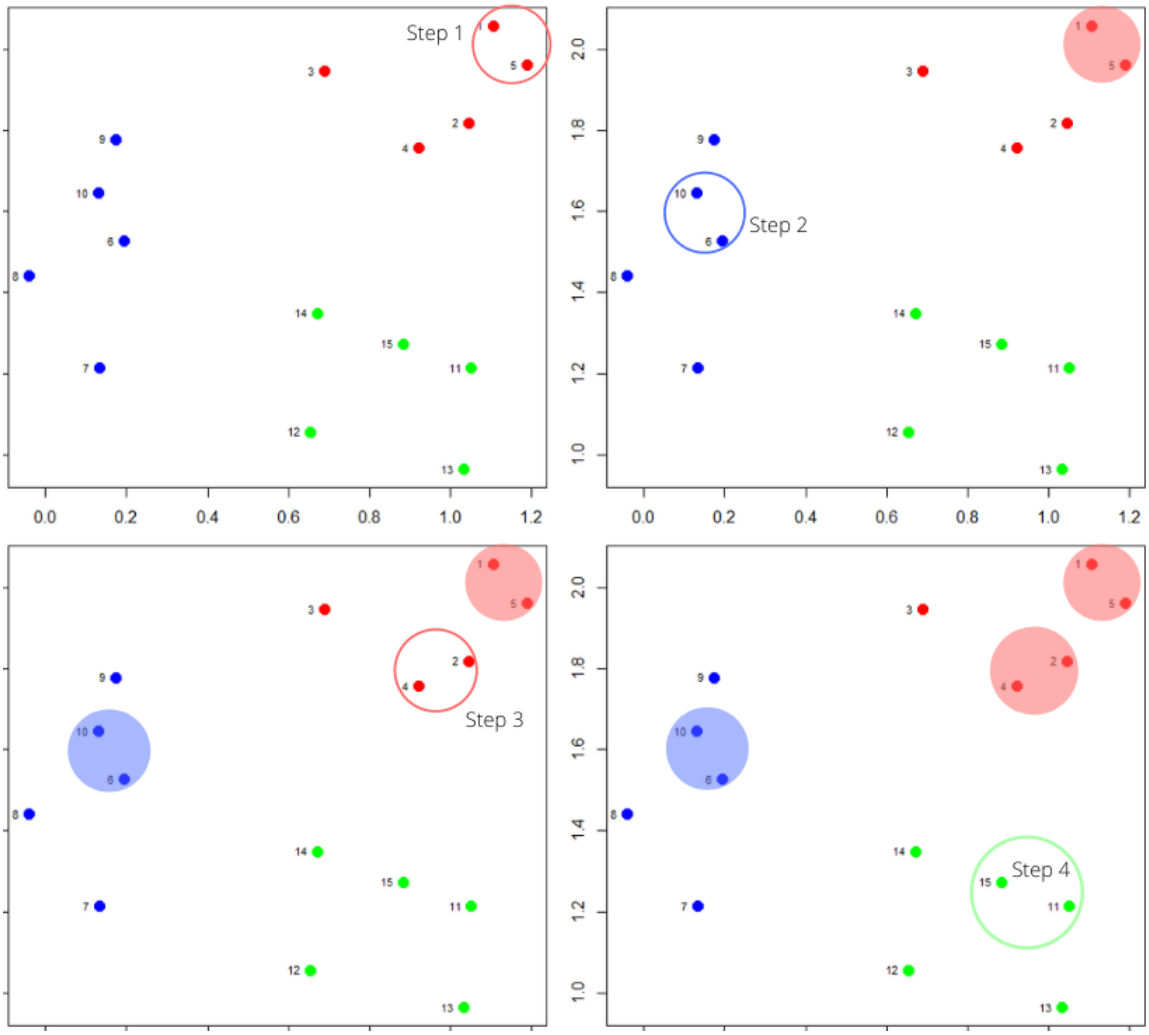


Figure 1.5: First four steps of the hierarchical clustering algorithm on the toy data set.

1.3.4. The choice of the number of clusters

The height at which the fusion of two groups happens, indicates how different these clusters are. If the fusion occurs at the bottom, the groups are very similar, on the other hand the higher the merge, the different the clusters are. Note that if two point are located near each other, this does not means anything in terms of similarity.
Looking at the dendogram obtained, two possible scenarios seem feasible. In Figure 1.6 are shown these two situations.

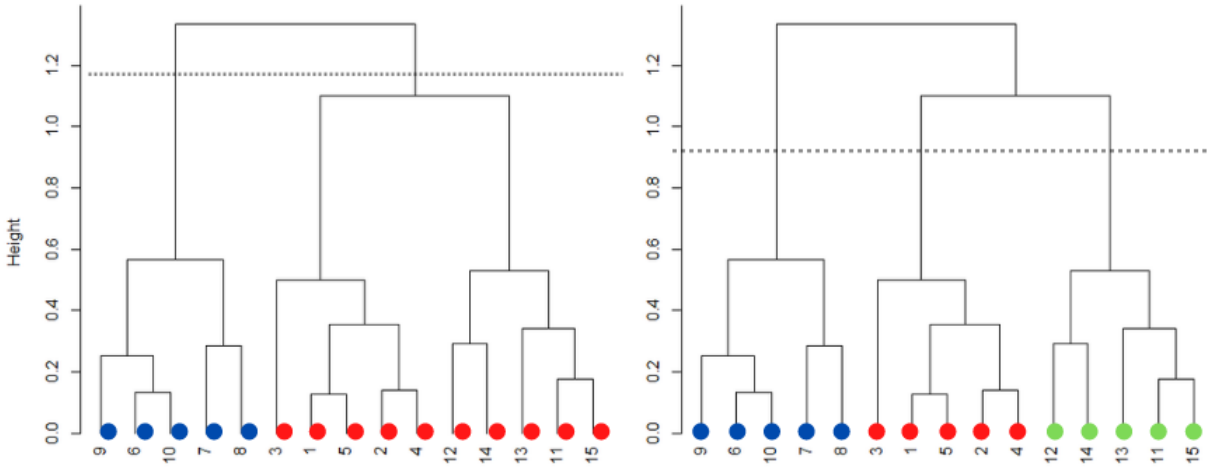


Figure 1.6: Clusters obtained cutting the dendogram at two different heights.

In the left panel, the cut, at height ≈ 1.2, result in two clusters. In the right panel, the cut occur at height ≈ 0.9 , gives the three expected groups. This practice is commonly used, one can look at the dendogram and select a reasonable number of clusters, based on the height of fusion. Note that this is not always possible, in these situations it is necessary to consider other algorithms.

Another necessary remark is the strong dependence of the output both from the distance and the linkage selected.
We considered just one dendogram, and this drove to some conclusions.
In Figure 1.7 we can see what we can obtain with the four linkages.

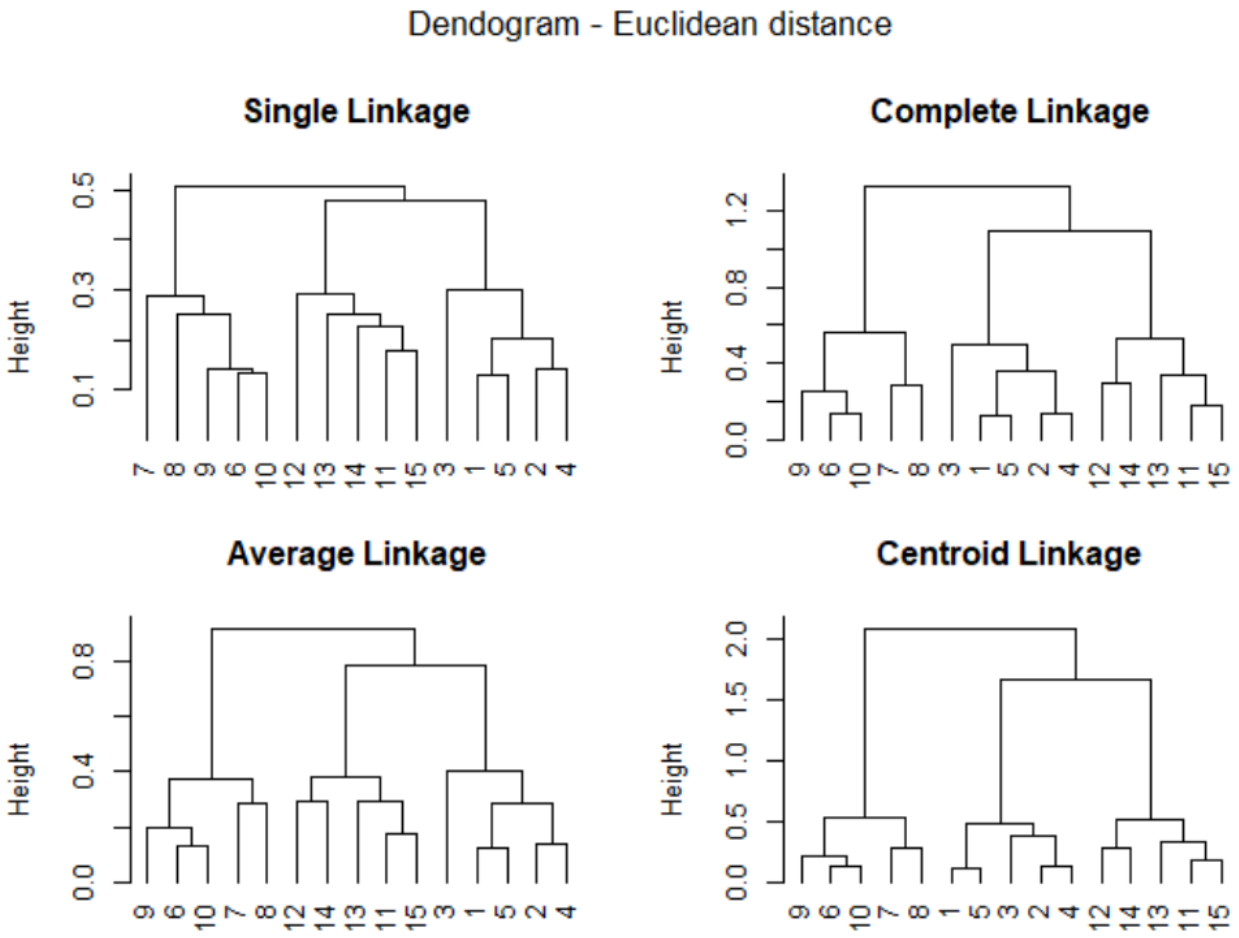


Figure 1.7: Dendograms obtained considering the four linkage presented.

In our simple scenario, there are no big differences, but just taking in consideration the dendogram obtained via single linkage one could immediately conclude that a feasible solution is three clusters.
The key point is stress that this algorithm can gives multiple solutions depending on the choice made, and a good practice is take different decisions and look at solution which result most useful.

2 | The structure of complex networks

2.1. Definitions and notations

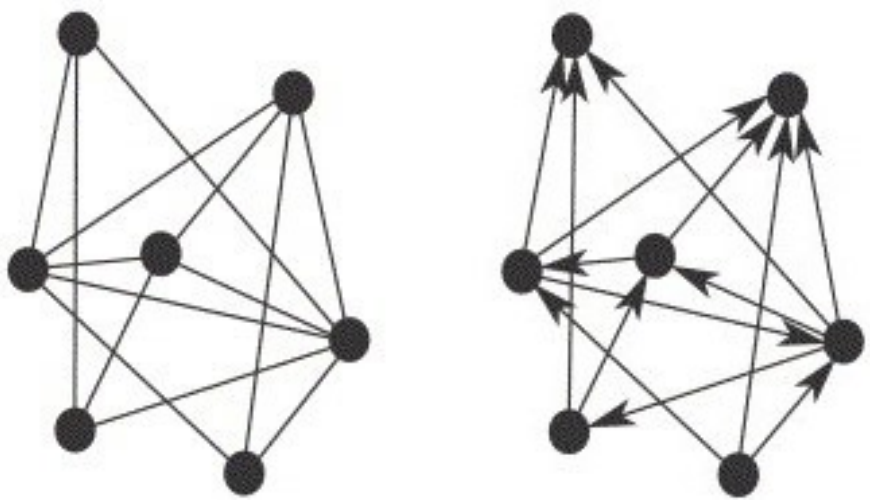
A undirected/direct) graph $\mathcal{G} = (\mathcal{N}, \mathcal{L})$ consists of two sets: the elements of $\mathcal{N} = \{n_1, n_2, ..., n_N\}$ are the nodes of the graph, while the elements of $\mathcal{L} = \{l_1, l_2, ..., l_K\}$ are its links.

In a undirected graph each of the link is defined for a couple of nodes i and j and it is denoted as l_{ij}.

Two nodes joined by a link are referred as adjacent.

In a directed graph, the order of the two nodes is important: $l_{ij} \neq l_{ji}$ in general.

The typical way to picture a graph is by drawing a dot for each node and joining two dots by a line if the corresponding link exists. Examples of undirected and directed graph are shown in Figure 2.1a and Figure 2.1b. In the directed graph, adjacent nodes are connected by arrows, indicating the direction of each link.



(a) Undirected graph. (b) Directed graph.

A graph $\mathcal{G} = (\mathcal{N}, \mathcal{L})$ can be completely described by the Adjacency matrix.

Definition 2.1. ***Adjacency matrix***
The Adjacency matrix is an N x N matrix: $A = (a_{ij})$ *such that*

$$a_{ij} = \begin{cases} 1, & \textit{if the link } l_{ij} \textit{ exists.} \\ 0, & \textit{otherwise} \end{cases}$$

Proposition 2.1. *The Adjacency matrix is symmetric if the graph is undirected, while is asymmetric if the graph is directed.*

2.1.1. Node degree

Definition 2.2. ***Node degree***
The degree k_i *of a node i is the numbers of links incident with the node, and is defined in terms of the Adjacency matrix as:*

$$k_i = \sum_{j \in \mathcal{N}} a_{ij}$$

If the graph is directed, the degree of the node has two components: the out-degree k_i^{out}, which refers to the number of outgoing links, and the in-degree k_i^{in}, which refers to the number of ingoing links. The total degree is then defined as $k_i = k_i^{in} + k_i^{out}$.

Definition 2.3. ***Laplacian matrix***
The Laplacian matrix is an N x N matrix: $L = diag\{k_1, k_2, ..., k_N\} - A$

Definition 2.4. ***Degree distribution***
The degree distribution P(k) is the probability that a node chosen uniformly at random has degree k, or equivalently, the fraction of nodes in the graph having degree k.

In the case of directed networks, one have to consider two distributions, $P(k^{in})$ and $P(k^{out})$.

Definition 2.5. ***The n-moment of P(k)***

$$\langle k^n \rangle = \sum_k k^n P(k)$$

2.1.2. Correlated networks

Definition 2.6. ***Conditional probability***
The conditional probability $P(k'|k)$ *is the probability that a link from a node with degree k*

points to a node of degree k'.

A network is correlated if the conditional probability depends on k, otherwise the network is uncorrelated.

Definition 2.7. ***Average nearest neighbors degree***
The average degree of the nearest neighbors of node i, can be defined as:

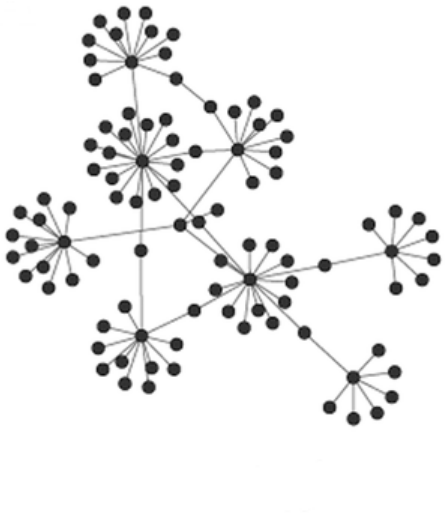
$$k_{nn,i} = \frac{1}{k_i} \sum_{j=1}^{N} a_{ij} k_j$$

Definition 2.8. ***Average nearest neighbors degree***
The average degree of the nearest neighbors of node with degree k, can be expressed in terms of the conditional probability as

$$k_{nn}(k) = \sum_{k'} k' P(k'|k)$$

Correlated graphs are classified as assortative if $k_{nn}(k)$ is an increasing function of k, and as disassortative when $k_{nn}(k)$ is a decreasing function of k. In assortative networks the nodes tend to connect to their connectivity peers, while in disassortative networks nodes with lower degree are more likely connected with highly connected ones. An example of this kind of behaviour is shown in Figure 2.1c and in Figure 2.1d.

(c) Assortative graph.



(d) Disassortative graph.

2.1.3. Distance and diameter

Definition 2.9. ***Distance***
The distance d_{ij} is the length, in terms of number of links, of the shortest path connecting node i to node j.

The matrix $\mathcal{D} = (d_{ij})$ represent all the shortest paths of a graph $\mathcal{G}$.

Definition 2.10. ***Diameter***
The diameter D is the maximum value of d_{ij}

$$D = \max_{ij} d_{ij}$$

Definition 2.11. ***Average distance***

$$d = \langle d_{ij} \rangle = \frac{1}{N(N-1)} \sum_{i,j:i \neq j} d_{ij}$$

A problem with last definition is that D diverges if there are disconnected components in the graph.Thus an alternative approach is to consider the harmonic mean of distances.

Definition 2.12. ***Efficiency***

$$E = \frac{1}{N(N-1)} \sum_{i,j \in \mathcal{N}: i \neq j} \frac{1}{d_{ij}}$$

2.1.4. Clustering coefficient

In a generic graph $\mathcal{G}$, clustering means the presence of a high number of triangles. This is a typical property of networks, where two individuals with a common friend are likely to know each other.

Definition 2.13. ***Local clustering coefficient***
The local clustering coefficient of node i is defined as:

$$C_i = \frac{e_i}{\frac{k_i(k_i-1)}{2}}$$

where e_i is the number of triangles connected to i.

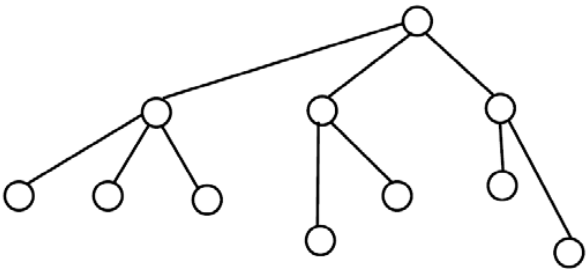
By definition, $0 \leq C_i \leq 1$. In summary C_i measures the network's local link density: The more densely interconnected the neighborhood of node i, the higher is its local clustering coefficient.

- $C_i = 0$: none of the neighbors of node i link to each other.
- $C_i = 1$: the neighbors of node i form a complete graph.

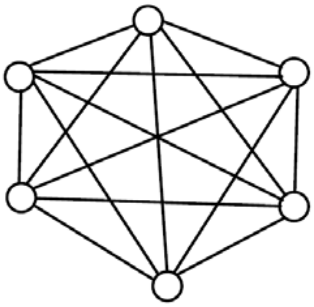
Definition 2.14. ***Clustering coefficient***
The clustering coefficient of a graph $\mathcal{G}$ is given by the average of c_i over all nodes

$$C = \langle c \rangle = \frac{1}{N} \sum_{i \in \mathcal{N}} c_i$$

By definition, $0 \leq C \leq 1$.Clearly $C = 1$ if and only if the network is globally coupled, i.e, every node in the network is linked to every other node.



(e) Tree network C = 0.



(f) Complete network C = 1.

2.1.5. Weighted networks

A undirected/directed weighted graph $\mathcal{G}_w = (\mathcal{N}, \mathcal{L}, \mathcal{W})$ consists of a set $\mathcal{N} = \{n_1, n_2, ..., n_N\}$ of nodes, a unordered/ordered set $\mathcal{L} = \{l_1, l_2, ..., l_K\}$ of links, and a set of weights $\mathcal{W} = \{w_1, w_2, ..., w_K\}$ that are real numbers attached to the links.
A undirected weighted graph can be drawn as in Figure 2.1, where the values $w_{i,j}$ reported on each link indicate the weights of the links, and are graphically represented by the link thicknesses. In the case of a directed one the usual arrow representation of the links holds.

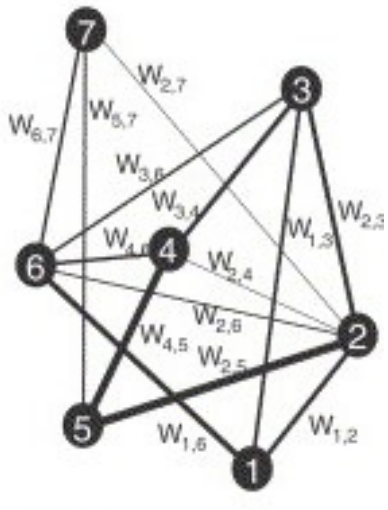


Figure 2.1: Weighted undirected graph

The metrics analyzed for unweighted graphs can be generalized for the weighted scenario [10].

Definition 2.15. ***Weights matrix***
The Weights matrix W is an N x N matrix, whose entry $w_{ij} > 0$ is the weight of the link connecting node i to node j, and $w_{ij} = 0$ if node i and node j are not connected.

Proposition 2.2. *The Weights matrix is symmetric if the graph is undirected, while is asymmetric if the graph is directed.*

Definition 2.16. ***Node strength***
In a weighted graph, the generalization of the degree k_i of a node i is the node strength s_i defined as:

$$s_i = \sum_{j \in \mathcal{N}} w_{ij}$$

If the graph is directed, the strenght of the node has two components: the out-strength:

$$s_i^{out} = \sum_{j} w_{ij}$$

which refers to the number of outgoing links, and the in-strength

$$s_i^{in} = \sum_{j} w_{ji}$$

which refers to the number of ingoing links. The total strength $s_i = s_i^{in} + s_i^{out}$.

Definition 2.17. ***Total network weight***
The total network weight is defined as:

$$S = \sum_{ij} w_{ij}$$

If the network is undirected, then $s_i = s_i^{in} = s_i^{out}$*, thus:*

$$S = \frac{1}{2} \sum_{ij} w_{ij}$$

Similarly to the degree distribution P(k), it is possible to define the strength distribution P(s), which measures the probability that a node has strength s.

Definition 2.18. ***Weighted average nearest neighbors degree***
The weighted average nearest neighbors degree of a node i, can be defined as:

$$k_{nn,i}^{w} = \frac{1}{s_i} \sum_{j \in \mathcal{N}} a_{ij} w_{ij} k_j$$

Such quantity is useful to charachterize the assortative/disassortative behavior in a weighted network:
when $k_{nn,i}^{w} > k_{nn,i}$, the links with larger weights are pointing to the neighbors with higher degree, and $k_{nn,i}^{w} < k_{nn,i}$ in the opposite scenario.

Definition 2.19. ***Weighted clustering coefficient***
The weighted clustering coefficient of a given node i, is defined as:

$$C_i^{w} = \frac{1}{s_i(k_i - 1)} \sum_{j,m} \frac{w_{ij} + w_{im}}{2} a_{ij} a_{jm} a_{mi}$$

2.2. Network models

In this section, we focus on the mathematical modelling of networks [31][15][25][12], discussing their construction procedure and highlighting significant properties.

2.2.1. Random networks: Erdös - Rényi

Starting with N disconnected nodes, random graphs are generated by connecting each couple of nodes with a probability $0 < p < 1$.

Proposition 2.3. ***Properties***
For large N, and fixed $\langle k \rangle$ we have:

- *The degree distribution is well approximated by a Poisson distribution:*

$$P(k) = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$$

- *Large networks have a relatively small average distance:*

$$d \approx \frac{log(N)}{log(\langle k \rangle)}$$

- *Large networks have vanishing clustering coefficient:*

$$C = p \approx \frac{\langle k \rangle}{N} \to 0$$

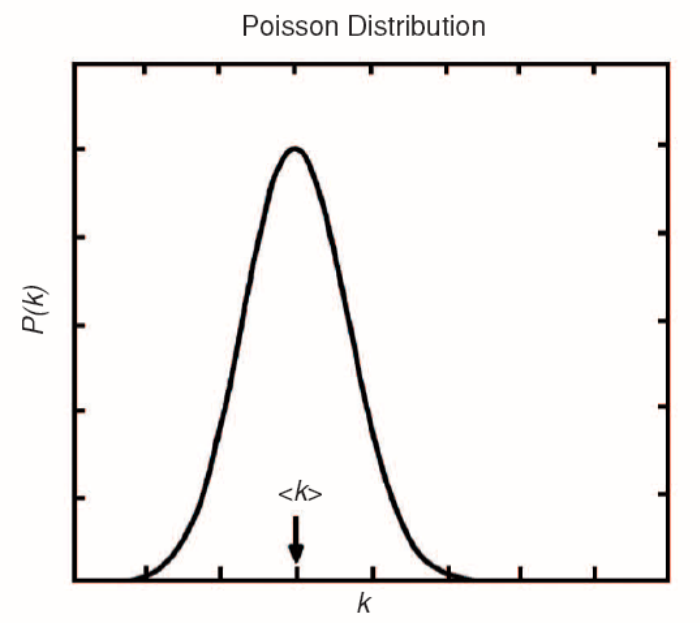


Figure 2.2: Poisson degree distribution

2.2.2. Scale-free networks: Barabási – Albert

Proposition 2.4. ***Scale-free networks***
Networks with a power law distribution of the form $P(k) \approx k^{-\alpha}$ *are called scale-free networks.*

The Barabási – Albert model is a model of network growth inspired to the formation of the World Wide Web, with the goal to reproduce the topological properties. The basic idea in the WWW, is that sites(nodes) with higher degrees acquire new links at higher rates than low-degree ones.
The Barabási – Albert graph is constructed as follow:
starting with m_0 isolated nodes, at each time step t, a new node j with $m \leq m_0$ links is added to the network.
The probability p that a link will connect j to an existing node i is proportional to the degree of i:

$$p = \frac{k_i}{\sum_h k_h}$$

Proposition 2.5. ***Properties***
In the limit $t \to +\infty$

- *The model produce the following power law distributon:*

$$P(k) \approx k^{-\gamma}, \gamma = 3$$

- *The network average degree is* $\langle k \rangle = 2m$*:*
- *The average distance in he model increase logarithmically with N:*

$$d \approx \frac{log(N)}{log(log(N))}$$

- *The clustering coefficient vanishes:*

$$C \approx \frac{log(N)^2}{N} \to 0$$

In Figure 2.3 are reported some examples of power law degree distribution in some large networks [12].

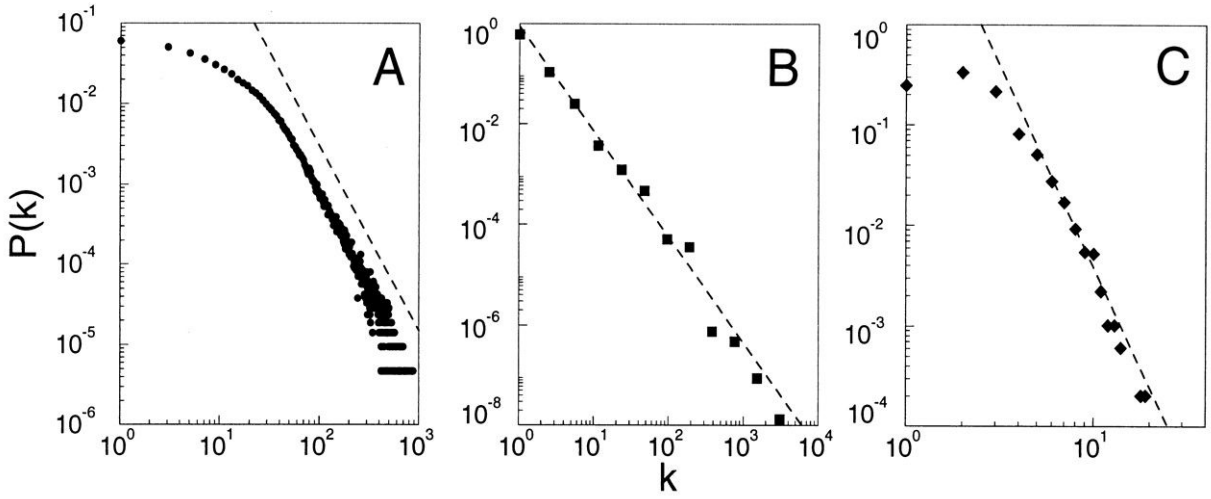


Figure 2.3: (**A**) The collaboration graph of movie actors. $\gamma \approx 2,3$. (**B**) WWW. $\gamma \approx 2,1$. (**C**) Electrical power grid of western US. $\gamma \approx 4$.

2.2.3. Small-world networks: Watts - Strogatz

Proposition 2.6. ***The small-world property***

The small-world property is mathematically characterized by an average distance d, that depends at most logarithmically on the network size N.

This feature characterize most of the real networks, whose study has pointed out the existence of shortcuts, links connecting distant nodes speeding up the communication.
The Watts and Strogatz model is a procedure to construct graphs, having both the small-world property and an high clustering coefficient. The starting point is a N node regular lattice, in which each node is connected to its 2m nearest neighbors, m right-neighbors and m left-neighbors.
The model is based on a rewiring procedure: for every node, each link is redirected to a randomly picked node with a probability p.

Proposition 2.7. ***Properties***
For intermediate values of p the obtained graph presents:

- *The degree distribution is concentrated around the average degree* $\langle k \rangle = 2m$.
- *The network has large clustering coefficient:*

$$C = \frac{3m-3}{4m-2}$$

- *The average distance decrease significantly, and passes from* $d \approx N$ *to* $d \approx log(N)$.

2.3. Node centralities

We now introduce a set of measures, which quantify the relevance of a node in a graph $\mathcal{G}$. These measures of node centrality are the degree(or the strength in case of a weighted graph $\mathcal{G}_w$), the betweenness, and the closeness.
The degree, trivially capture the importance of a node i,by counting the numbers of its k_i neighbors.
However more relevant measures are the betweennes and the closeness.
The betweennes quantify the most relevant nodes, as the ones who have the control on the flow of information between most others.

Definition 2.20. ***Betweennes***
The betweenness b_i of a node i, is defined as:

$$b_i = \sum_{j,k \in \mathcal{N}: j \neq k} \frac{n_{jk}(i)}{n_{jk}}$$

where n_{jk} is the number of shortest paths connecting j and k, and $n_{jk}(i)$ is the number of shortest paths connecting j and k passing through i.

A node is central if, on average, it is closed, in terms of shortest paths, to all other nodes. This is quantified by the closeness.

Definition 2.21. ***Closeness***
The closeness of a node is defined as the inverse of the average distance from all other nodes:

$$c_i = \frac{n-1}{\sum_j d_{ij}}$$

2.4. Community detection

Graphs representing real systems are not regular, however they display big inhomogeneities, revealing a high level of order and organization. The distribution of links, can be, even at local level, highly concentrated within groups of nodes, and lower between these groups. This feature is called community structure [17].
Communities are groups of nodes which probably share properties or have similar roles in the network.

Definition 2.22. ***Sparse graphs***
A graph is sparse if the number of links L is of the order of the number of nodes N.

The identification of structural clusters is possible only if graphs are sparse. Otherwise if $L \gg N$, the link distribution among the nodes is too homogeneous for communities to make sense.
In most cases, communities are algorithmically defined without a precise a priori definition. However before introducing some algorithms, the idea at basis of these is the following [18]:
Let us begin with a subgraph $\mathcal{C}$ of a graph $\mathcal{G}$, with N_c nodes.

Definition 2.23. ***Intra-cluster density***
The intra-cluster density $\delta_{int}(\mathcal{C})$ is the ratio between the number of links connecting the nodes of $\mathcal{C}$, L_C^{int}, over all possible internal links:

$$\delta_{int}(\mathcal{C}) = 2\frac{L_C^{int}}{N_c(N_c-1)}$$

Definition 2.24. ***Inter-cluster density***
The inter-cluster densisty $\delta_{ext}(\mathcal{C})$ is the ratio between the number of links, L_C^{ext}, connecting the nodes of $\mathcal{C}$ to the rest of the graph $\mathcal{G}$ and the maximum number of inter-cluster links possible:

$$\delta_{ext}(\mathcal{C}) = \frac{L_C^{ext}}{N_c(N-N_c)}$$

For $\mathcal{C}$ to be a community we expect $\delta_{int}(\mathcal{C})$ to be larger than the average link density of $\mathcal{G}$, i.e, the ratio between the number of link L and the maximum number of possible links N(N-1)/2. At the same time, $\delta_{ext}(\mathcal{C})$ has to be smaller.
Searching for the best trade-off between a small $\delta_{ext}(\mathcal{C})$ and a large $\delta_{int}(\mathcal{C})$ is the goal of most of clustering algorithms.

2.4.1. The method of Modularity optimization

Given a network $\mathcal{G} = (\mathcal{N}, \mathcal{L})$, suppose we want to determine whether there exist any division of its nodes into non-overlapping communities, and furthermore, these communities may be of any size [26].
Two communities case
Let us begin, considering that any good division of $\mathcal{G}$ exists into just two communities: $\mathcal{C}_1$ and $\mathcal{C}_2$. A good community is one in which there are fewer than expected links between communities. If the number of link between two groups is only what one would expect on the basis of random chance, then few thoughtful observer would claim a meaningful community structure. On the other hand, if the number of links between two groups is

less than we expect by chance, or the number within groups is significantly more, then a partition could be relevant.
This idea is quantified by the modularity:
For a division of $\mathcal{G}$ into $\mathcal{C}_1$ and $\mathcal{C}_2$, let:

$$s_i = \begin{cases} 1, & \text{if i } \in \text{C}_1 \\ -1, & \text{if i } \in \text{C}_2 \end{cases} \quad i = 1, ..., N$$

The expected number of links between two nodes i and j, if nodes are placed at random is $k_i k_j / 2L$

Definition 2.25. ***Modularity***
The modularity $\mathcal{Q}$ is the number of links falling within a group minus the expected number in an equivalent network with links placed at random:

$$\mathcal{Q} = \frac{1}{4L} \sum_{ij} \left(a_{ij} - \frac{k_i k_j}{2L} \right) s_i s_j$$

Definition 2.26. ***Modularity matrix***
By the previous definition of modularity is possible to define the real and symmetric matrix, called the modularity matrix $B = (b_{ij})$ as:

$$b_{ij} = a_{ij} - \frac{k_i k_j}{2L}$$

By the way, it is convenient to rewrite the modularity $\mathcal{Q}$ in a vectorial form: set $\mathbf{s}$ the column vector whose element are s_i, and considering the modularity matrix B then:

$$\mathcal{Q} = \frac{1}{4L} \mathbf{s}^T B \mathbf{s}$$

We can proceed by writing $\mathbf{s}$ as a linear combination of the normalized eigenvectors u_i of B: $\mathbf{s} = \sum_i \lambda_i u_i$. Then:

$$\mathcal{Q} = \frac{1}{4L} \sum_i \lambda_i u_i^T B \sum_j \lambda_j u_j = \frac{1}{4L} \sum_i \left(u_i^T \mathbf{s} \right)^2 \beta_i$$

where β_i is the eigenvalue of B corresponding to u_i. Note that each row sum to 0, thus it always admits $(1, 1, ..., 1)$ as eigenvector associated to 0 as eigenvalue.
Suppose that the eigenvalues are labelled in decreasing order: $\beta_1 \geq \beta_2 \geq ... \geq \beta_N$.
We want to maximize the modularity by choosing the value of the vector $\mathbf{s}$, considering

the restriction that $s_i = \pm 1$. The problem reduces to maximizing the dot product $u_1^T \mathbf{s}$. The maximum is reached by setting:

$$s_i = \begin{cases} 1, & \text{if } (u_1)_i > 0 \\ -1, & \text{if } (u_1)_i < 0 \end{cases} i = 1, ..., N$$

This algorithm provide the separation into the two cluster by computing the dominant eigenvalue of B and divide the nodes according to the sign of the element in the vector. Notice that the size of the communities are not specified, and a network is indivisible if B has no positive eigenvalues.

More than two communities case

Networks, however, in many cases contain more than two communities, thus the purpose now is to generalize the method discussed in the previous paragraph, in order to divide the graph into larger numbers of parts. The approach is to consider the additional contribution $\Delta\mathcal{Q}$ upon further dividing a group g of size N_g in two:

$$\Delta\mathcal{Q} = \frac{1}{4L}\mathbf{s}^T B^g \mathbf{s}$$

where $B^g = (b_{ij}^g)$ is the $N_g \times N_g$ matrix with element indexed by the labels i,j of nodes belonging to group g:

$$b_{ij}^g = b_{ij} - \delta_{ij} \sum_{k \in g} b_{ik}$$

where δ_{ij} is the Kronecker δ.

Now we have all the elements to proceed as in the two communities scenario, considering the spectral approach in order to maximize $\Delta\mathcal{Q}$. Note that the rows and columns of B^g sum to zero, and $\Delta\mathcal{Q}$ is null if group g is undivided. This happens when there are no positive eigenvalues to B^g.

Proposition 2.8. *The absence of positive eigenvalues to the matrix B^g is a sufficient condition for indivisibility of the group g.*

The algorithm is the following: consider the modularity matrix B for the graph $\mathcal{G}$ and find its dominant eigenvaues and the corresponding eigenvector.Divide the network into two parts according to the sign of the elements of this vector.

Repeat the process for each of the parts using the generalize modularity matrix B^g.

If at any stage the proposed division makes a zero or negative contribution to the modularity , leave the corresponding sub-graph undivided.

The algorithm ends when the entire network has been decomposed into indivisible sub-

graphs.

2.4.2. Finding communities by means of random walkers

The modularity index maximization, in some cases, due to intrinsic limitations, it does not produce a significant partition. And even when it does, it quantifies the quality of the whole partition and not of each community.
We take into consideration a definition of community which is based on a quality threshold $0 < \alpha < 1$ [28].

Methods

A N-state Markov chain $\pi_{t+1} = \pi_t P$, with $\pi_t = (\pi_1, \pi_2, ..., \pi_N)_t$ can be associated to the network, row-normalizing the weight matrix W:

Definition 2.27. ***Transition probability***
The transition probability matrix $P = (p_{ij})$, whose entrance is the transition probability that a random walker in node i goes to node j, is defined as:

$$p_{ij} = \frac{w_{ij}}{\sum_j w_{ij}}$$

$$0 \leq p_{ij} \leq 1$$

$$\sum_j p_{ij} = 1$$

Proposition 2.9. *P is irreducible if the network is connected.*

If P is irreducible , than the equation $\pi = \pi P$ has a unique solution $\pi = (\pi_1, \pi_2, ..., \pi_N)$, strictly positive, i.e, $\pi_i > 0 \quad \forall i$.

Proposition 2.10. *For undirected networks, the stationary Markov chain state probability distribution is:*

$$\pi = \frac{1}{2S}(s_1, s_2, ..., s_N)$$

Let us consider $\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_q$, q disjointed sub-networks of $\mathcal{G}$, which are the candidate communities. The description of the dynamics of the random walker is given by a lumped Markov chain, defined by the $q \times q$ row-stochastic matrix:

$$U = [diag(\pi H)]^{-1} H^T diag(\pi) P H$$

where $H = (h_{ic})$ is an $N \times q$ matrix, with $h_{ic} = 1$ if and only if the node i $\in \mathcal{C}_c$.

Definition 2.28. ***Persistence probability***

The persistence probability of a community $\mathcal{C}_c$ is the diagonal element of U, u_{cc}.

Definition 2.29. *α-community*
Given a value $0 < \alpha < 1$, $\mathcal{C}_c$ is an α - community if $u_{cc} \geq \alpha$. Thus α act as a quality threshold.

Definition 2.30. *α-partition*
A partition $\mathcal{P}_q = (\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_q)$, is an α-partition, if it is composed by α-communities, i.e, $u_{cc} \geq \alpha \quad \forall c = 1, 2, ..., q$.

Testing the quality of a partition
Testing the quality of a given partition, no matter how obtained, become extremely easy by means of persistence probability.
Fixing the quality value α, for instance a typical value is $\alpha = 0.5$, then we have just to verify if the considered partition is whether α or not.
It is possible to derive an explicit expression of the persistence probability u_{cc} of a community $\mathcal{C}_c$:

$$u_{cc} = \frac{\sum_{i,j \in \mathcal{C}_c} \pi_i p_{ij}}{\sum_{i \in \mathcal{C}_c} \pi_i}$$

In the case of undirected network, we obtain:

$$u_{cc} = \frac{\sum_{i,j \in \mathcal{C}_c} w_{ij}}{\sum_{i \in \mathcal{C}_c} s_i}$$

Finding α-partitions
The starting point is to define the level for the quality parameter α.
The formulation of the problem of community detection is the following:

$$\begin{aligned} \max_{\mathcal{P}_q \in \mathcal{P}} \quad & q \\ \text{s.t.} \quad & u_{cc} \geq \alpha \quad \text{c} = 1,2,...,\text{q} \end{aligned}$$

where $\mathcal{P}$ is the set of all partitions.
Notice that the set of admissible set of problem is not empty for any given α, since $\mathcal{P}_1 = \mathcal{G}$ has $u_{11} = 1$. Moreover, in general the optimal solution is not unique.
We consider an heuristic approach for finding a sub-optimal solution.
First, restrict the optimization to a sub-set $\mathcal{P}^* \subset \mathcal{P}$, obtained by any algorithm.
The problem solution is the α-partition in $\mathcal{P}^*$ with the largest q.
The approach is the following:

- Generate a collection of meaningful partitions.

- Compare their quality.
- Select the preferred partition, implicitly fixing the quality α a posteriori.

Cluster analysis can be used to produce candidate partitions. In order to proceed we have to define a similarity among each couple of nodes in the graph.
Consider a large number M of repetitions of a random walker started from i. For each repetition, the probability that the random walker is in j at step t, is $(P^t)_{ij}$.
If the length of the random walk is T, a possible symmetric similarity is the following:

$$\sigma_{ij} = \sum_{i=1}^{T} \left[(P^t)_{ij} + (P^t)_{ji} \right]$$

Then we define the distance between two nodes, i and j, by normalizing the similarity metric σ:

$$d_{ij} = 1 - \frac{\sigma_{ij} - \min \sigma_{ij}}{\max \sigma_{ij} - \min \sigma_{ij}}$$

The choice of the time horizon T, can become crucial. Cluster analysis drives to a different dendogram for each T. An effective selection of T can be empirically obtained by maximizing the cophenetic correlation coefficient, i.e, the linear correlation between the distance d_{ij} and the cophenetic distance c_{ij}. To summarize the procedure:

- Apply cluster analysis for T that maximizes the cophenetic correlation coefficient.
- Top-down section of the associated dendogram identifies the sequence of partitions $\mathcal{P}_1, \mathcal{P}_2, ...$, with an increasing number of communities.
- For each $\mathcal{P}_q$, compute the lumped Markov matrix U, and then compare the diagonal terms.
- Choose the value q, the maximum which satisfies the quality threshold α.

2.5. Link prediction

Given an undirected network $\mathcal{G} = (\mathcal{N}, \mathcal{L})$, with N nodes and L links.
Assume that multiple links and self-connection are not allowed.
Let $\mathcal{U}$ be the set containing all possible N(N-1)/2 links. Then the set of non-existing links is $\mathcal{U} \setminus \mathcal{L}$. The purpose of link prediction is to detect the actual missing links in the network $\mathcal{G}$ evaluating $\mathcal{U} \setminus \mathcal{L}$.

2.5.1. Similarity-based algorithms

A link prediction algorithm outputs a list of non-observed links, giving each one a score s_{ij}. The higher the score, the higher the likelihood the link exists.
One of the simplest framework of link prediction methods is the similarity-based algorithm [23]. In particular, we will focus on a group of similarity indices, based only on the network structure.

- **Common Neighbors:**
  For a node i, let B_i be the set of its neighbors. Two nodes i,j are more likely to be connected by a link if they have many common neighbors. The simplest score is obtained by counting the number of common neighbors between two nodes, or equivalently considering the number of path of length two connecting them:

$$s_{ij} = |B_i \cap B_j| = (A^2)_{ij}$$

- **Jaccard Index:**
  This index is based on the normalized number of common neighbors:

$$s_{ij} = \frac{|B_i \cap B_j|}{|B_i \cup B_j|}$$

- **Resource Allocation Index:**
  Consider a pair of nodes, i and j, which are not directly connected. Node i can send some resource to j through their common neighbors. Assuming a transmission of a single unit of resource, and distribute it to all its neighbors:

$$s_{ij} = \sum_{l \in (B_i \cap B_j)} \frac{1}{k_l}$$

- **Adamic-Adar Index:**
  This index, based on common neighbors, assign the less connected neighbors more weight:

$$s_{ij} = \sum_{l \in (B_i \cap B_j)} \frac{1}{log(k_l)}$$

- **Preferential Attachment Index:**
  The probability that a new link connecting node i and node j, is proportional to the

product of their degrees:

$$s_{ij} = k_i \cdot k_j$$

3 | Database report: Social pillar

3.1. ESG and Social pillar

The acronym ESG refers to the incorporation of Environmental, Social and Governance issues into investment decisions.
ESG factors are typically associated with non-financial data, such as the environmental (e.g., CO_2 emissions), social (e.g., worker protection) and governance (e.g., board composition) impact of a given company.
In recent years, investments that include ESG factors represent an increasingly large share of global investments, and have been estimated to be in the tens of trillions of USD in AUM.[24]
This growth is also due to the realisation that ESG factors are a source of transitional risk, due to the costs and possible impacts of new climate policies, and physical risk, stemming from the increase in extreme weather events and structural climate variations, and thus affecting the value of companies in the long run.
This report will examine data related to the social pillar.
The social pillar includes aspects related to gender policies, human rights protection, and workplace standards.
The factors that will be analysed concern aspects that influence employee satisfaction, such as training and remuneration.[14]

3.2. Database costruction

Legislative Decree No. 254/2016, introduced the obligation to publish a non-financial statement, the Dnf, for Italian companies listed on an Italian or EU regulated market.[1]
The framework to be followed in drafting this document is the GRI Susteinability reporting standards.[3]
The GRI indices considered, related to the social pillar, will be presented in detail in a later section.
The data collection involves the search for numerical values corresponding to these GRI

indices. The values were collected, if available, from the published "Dnf" of 164 Italian listed companies belonging to the FTSE All Share, taking 2020 as the reporting year. The list of companies considered follows:

1. A2A SPA
2. ABITARE IN
3. ACEA SPA
4. AEFFE SPA
5. AEROPORTO GUGLIELMO MARCONI DI BOLOGNA SPA
6. ALKEMY SPA
7. AMBIENTHESIS
8. AMPLIFON
9. ANIMA HOLDING
10. AQUAFIL
11. ASCOPIAVE SPA
12. ATLANTIA SPA
13. AUTOGRILL SPA
14. AVIO SPA
15. AZIMUT HOLDING
16. BANCA CARIGE
17. BANCA FARMAFACTORING SPA
18. BANCA GENERALI SPA
19. BANCA IFIS SPA
20. BANCA MEDIOLANUM SPA
21. BANCA MONTE DEI PASCHI DI SIENA SPA
22. BANCA POPOLARE DELL'EMILIA ROMAGNA, SOCIETA' COOPERATIVA
23. BANCA POPOLARE DI SONDRIO
24. BANCO BPM SPA
25. BASIC NET SPA
26. BE THINK, SOLVE, EXECUTE SPA
27. BEGHELLI
28. BF SPA
29. BIALETTI INDUSTRIE
30. BIESSE SPA
31. BREMBO
32. BRUNELLO CUCINELLI SPA
33. BUZZI UNICEM SPA
34. CAIRO COMMUNICATION SPA
35. CALTAGIRONE SPA
36. CAMPARI
37. CAREL INDUSTRIES SPA
38. CELLULARLINE SPA
39. CEMBRE SPA
40. CEMENTIR HOLDING
41. CENTRALE DEL LATTE D'ITALIA
42. CERVED GROUP SPA
43. CNH INDUSTRIAL

44. COFIDE - GRUPPO DE BENEDETTI SPA (CIR)
45. COIMA RES
46. CREDITO EMILIANO SPA
47. CSP INTERNATIONAL
48. D'AMICO
49. DANIELI
50. DATALOGIC
51. DE LONGHI
52. DIASORIN
53. DOVALUE
54. EDISON RSP
55. EL. EN. SPA
56. ELICA
57. EMAK
58. ENAV
59. ENEL
60. ENI
61. ERG SPA
62. ESPRINET
63. EXPRIVIA
64. FALCK RENEWABLES
65. FERRARI
66. FIERA MILANO
67. FILA
68. FINCANTIERI SPA
69. FINECOBANK
70. FNM SPA
71. GAROFALO HEALTHCARE SPA
72. GEFRAN SPA
73. GEOX
74. GPI SPA
75. GRUPPO MUTUIONLINE SPA
76. GVS
77. HERA SPA
78. IGD - SIIQ
79. IL SOLE 24 ORE
80. ILLIMITY BANK
81. IMMSI SPA
82. INDEL B
83. INTERPUMP GROUP SPA
84. INTESA SANPAOLO SPA
85. INWIT
86. IRCE
87. IREN SPA
88. ITALGAS SPA
89. ITALIAN EXHIBITION GROUP
90. ITALMOBILIARE
91. IVS GROUP
92. LA DORIA SPA
93. LANDI RENZO SPA
94. LEONARDO SPA
95. LUVE SPA

96. MAIRE TECNIMONT SPA
97. MARR SPA
98. MEDIASET SPA
99. MEDIOBANCA - BANCA DI CREDITO FINANZIARIO SPA
100. MONCLER SPA
101. MONDADORI EDITORE SPA
102. MONRIF SPA
103. NEODECORTECH SPA
104. NEWLAT FOOD SPA
105. NEXI SPA
106. OPENJOBMETIS SPA
107. ORSERO SPA
108. OVS SPA
109. PIAGGIO C SPA
110. PININFARINA SPA
111. PIOVAN SPA
112. PIQUADRO SPA
113. PIRELLI SPA
114. PLC SPA
115. POSTE ITALIANE SPA
116. PRIMA INDUSTRIE SPA
117. PRYSMIAN SPA
118. RAI WAY SPA
119. RATTI SPA
120. RCS MEDIAGROUP SPA
121. RECORDATI INDUSTRIA CHIMICA E FARMACEUTICA SPA
122. RENO DE MEDICI SPA
123. REPLY SPA
124. RETELIT SPA
125. SABAF SPA
126. SAES GETTERS SPA
127. SAFILO GROUP SPA
128. SALCEF GROUP SPA
129. SALVATORE FERRAGAMO SPA
130. SANLORENZO SPA
131. SARAS SPA RAFFINERIE SARDE
132. SECO
133. SERI INDUSTRIAL
134. SERVIZI ITALIA SPA
135. SESA SPA
136. SIT SPA
137. SNAM RETE GAS SPA
138. SOCIETA' CATTOLICA DI ASSICURAZIONE SOCIETA' COOPERATIVA
139. SOGEFI SPA
140. SOL SPA
141. SOMEC
142. STELLANTIS
143. STMICROELECTRONICS
144. TAMBURI INVESTMENT PARTNERS SPA

145. TAS
146. TECHNOGYM SPA
147. TELECOM ITALIA
148. TENARIS
149. TERNA RETE ELETTRICA NAZIONALE SPA
150. TESMEC
151. TINEXTA
152. TISCALI
153. TOD'S
154. TREVI FIN INDUSTRIALE SPA
155. TXT E-SOLUTIONS SPA
156. UNICREDIT SPA
157. UNIEURO SPA
158. UNIPOL GRUPPO FINANZIARIO SPA
159. UNIPOLSAI SPA
160. VALSOIA SPA
161. WEBUILD SPA
162. WIIT SPA
163. ZIGNAGO SPA - INDUSTRIE ZIGNAGO S. MARGHERITA
164. ZUCCHI SPA

For all societies was also considered the sector they belong to, following GICS. In Figure 3.1 an overview of the distribution over the 11 sectors.

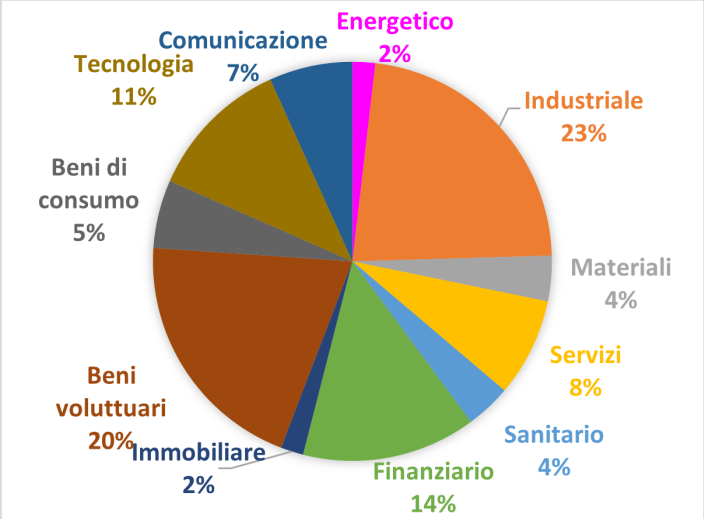


Figure 3.1: Sectors

Moreover, was taken into account if a company belongs or not to the MIB ESG index, considering the composition on the 28 April 2022.[4]
The companies analized, which are in the index are 38.

3.2.1. GRI standard indices

In this section will be presented the GRI indeces which were collected in the sustainability reports(Dnf).

Guidelines: The age groups considered are the following:

- under 30 years old.
- 30 - 50 years old.
- over 50 years old.

The emloyee categories considered are the following:

- dirigenti.
- quadri.
- impiegati.
- operai.
- 401-1. New employee hires and employee turnover.[5]
  - 401-1 a. Total number of new employee hires during the reporting period, by age group.
  - 401-1 b. Total number of new employee hires during the reporting period, by gender.
  - 401-1c. Rate of employee turnover during the reporting period, by age group.
  - 401-1d. Rate of employee turnover during the reporting period, by gender.

$$\text{rate of turnover} = \frac{\text{employess who leave the society}}{\text{total number of employees}}\%$$

- 403-9. Work-related injuries.[6]
  - 403-9 a.The number of fatalities as a result of work-related injury.
  - 403-9 b. Rate of recordable work-related injuries.

$$\text{rate of injuries} = \frac{\text{number of recordable work-related injuries}}{\text{number of hours worked}} \cdot 1000000$$

- 404-1. Average hours of training per year per employee.[7]

  - 404-1 a. Average hours of training that the organization's employees have undertaken during the reporting period, by gender.
  - 401-1 b. Average hours of training that the organization's employees have undertaken during the reporting period, by employee category.

- 405-1. Diversity of governance bodies and employees. [8]
  - 405-1 a. Percentage of women within the CdA(tradurre).
  - 405-1 b. Percentage of individuals within the CdA(board of directors?) by age groups.
  - 405-1 c. Percentage of women by employee category.
  - 405-1 d. Percentage of employees per employee category, by age group.
- 405-2. Ratio of the remuneration of women to men for each employee category.

3.3. Database analysis

General comments

The performed analysis are just a preliminary graphical overview on data distributions. No tests were performed to asses the statistical evidence on the observations which will be done.

Preliminary considerations: %NA and discrepancies from standards

Before going into detail on the individual types of data, it is worth highlighting the presence of NA and discrepancies with the data requested. NA may be due to a lack of data or to a deformity.

Since numerous discrepancies are present, special columns have been created to report the information anyway.

The main dissimilarities concern the categories of workers. Many companies have reported the figure considering either merged categories (e.g. Dirigenti e Quadri, etc.) or additional occupational categories (e.g. White collars and Blue collars, etc.) sometimes typical ad hoc companies (e.g. Giornalisti). Therefore, headings 404 and 405 have a high number of missing values for these reasons.

Some considerations will be made, where possible, on these discrepancies.

401-1

- 401-1 a.

	NA (%)	Mean	Min	Max	Std
<30	18.3	467.2	0	11775	1331.5
30-50	19.5	432.1	0	8650	1083.3
>50	19.5	81.7	0	1640	214.1

Table 3.1: 401-1a

The number of hired employees is higher for under 30 and 30-50. Moreover, firms hiring more are in the following sectors: beni voluttuari e industriale. Societies belonging to the MIB ESG index hire more personal on average.

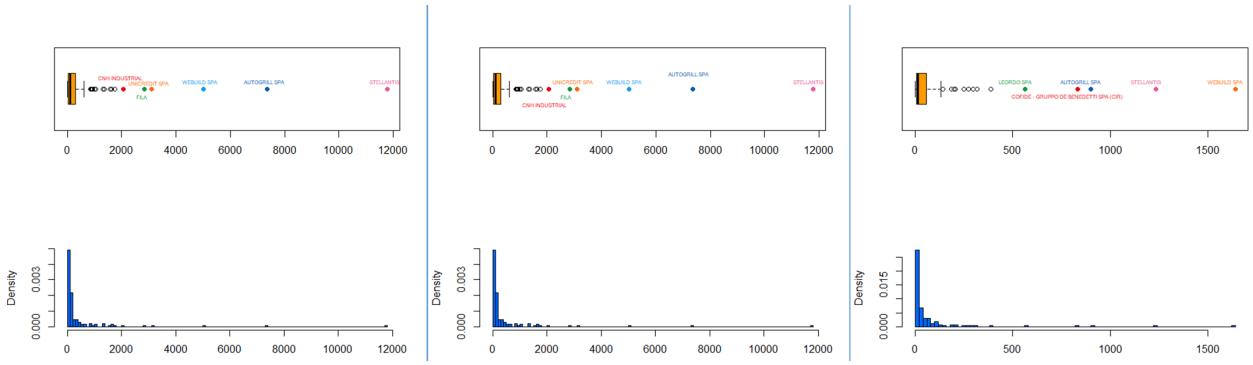


Figure 3.2: 401-1a: Left: <30; Center:30-50; Right:>50

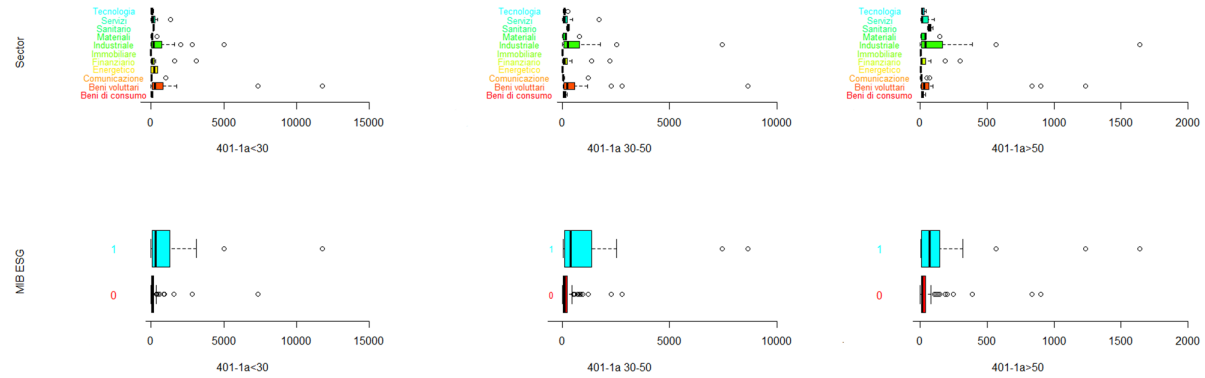


Figure 3.3: 401-1a. Sectors & MIB ESG boxplots

- 401-1 b.

	NA (%)	Mean	Min	Max	Std
D	14.6	383.7	0	7045	966.2
U	15.2	585.4	0	14614	1552.6

Table 3.2: 401-1b

Men are hired more than women, about 50% more. At a hiring level, the patthern seems to underline a potential discrimination by gender.

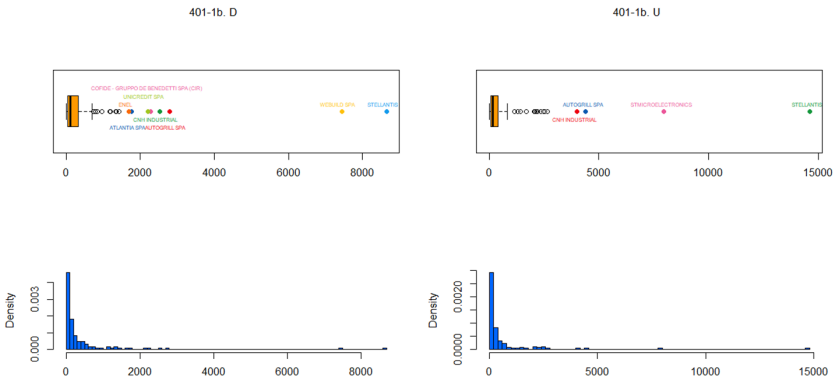


Figure 3.4: 401-1b: Left: D; Right:U

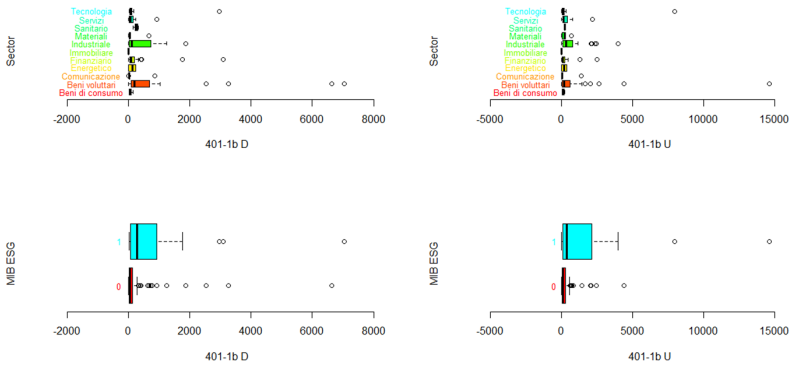


Figure 3.5: 401-1b. Sectors & MIB ESG boxplots

- 401-1 c.

	NA (%)	Mean	Min	Max	Std
<30	20.1	5.0	0	70.5	8.1
30-50	20.1	6.5	0	43.54	6.4
>50	20.1	3.5	0	24.1	3.3

Table 3.3: 401-1c

The rate of tournover, on average, is higher for under 30 and 30-50. This can have two possible interpretations: on one side it can indicate levels of uncertainty and dissatisfaction among employees, on the other side it signal a fundamental change in the structure of an organization's core. A comparison with the number of hiring could be useful to investigate more these data.

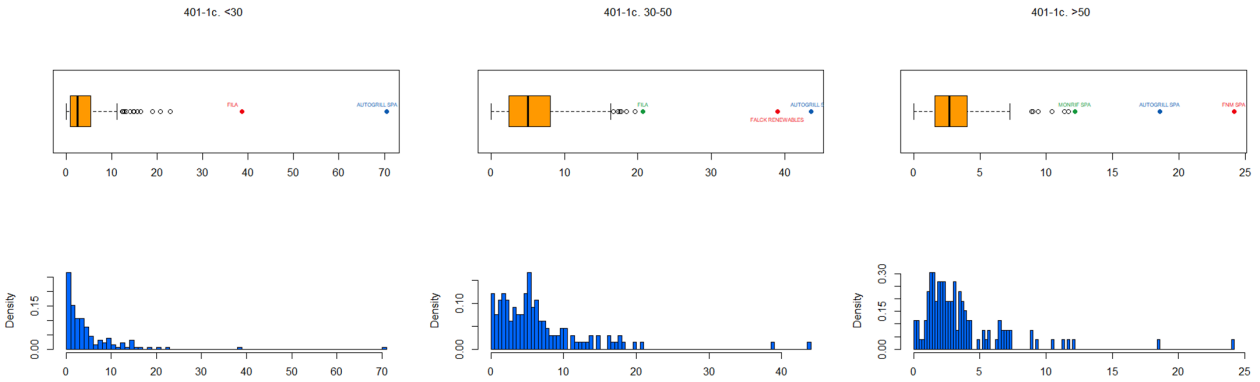


Figure 3.6: 401-1c.

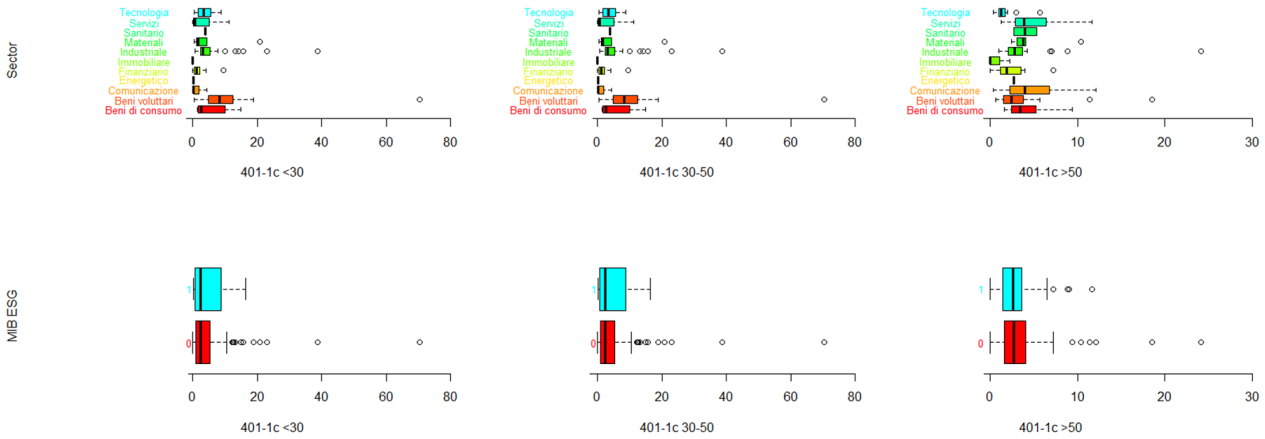


Figure 3.7: 401-1c. Sectors & MIB ESG boxplots

- 401-1 d.

	NA (%)	Mean	Min	Max	Std
D	23.8	6.4	0	78.5	9.2
U	19.5	8.9	0	54.3	7.9

Table 3.4: 401-1d

There is no significant difference for the rate of turnover by gender.

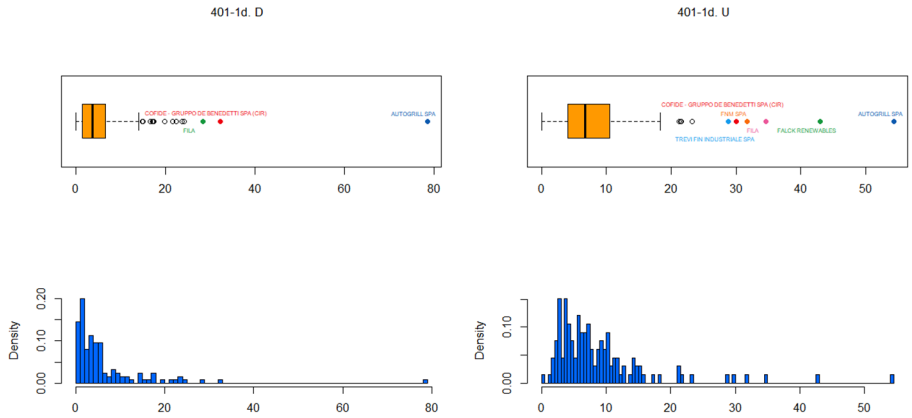


Figure 3.8: 401-1d.

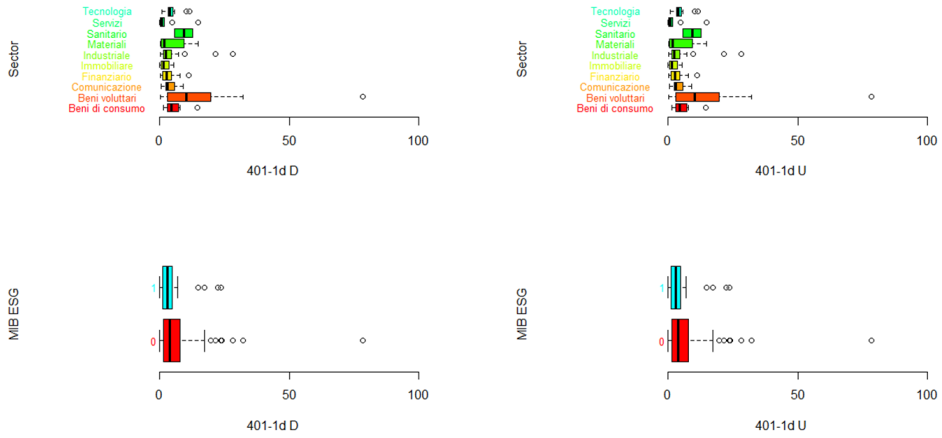


Figure 3.9: 401-1d. Sectors & MIB ESG boxplots

403-9

- 403-9a.

	NA (%)	Mean	Min	Max	Std
N Inf	19.2	0.3	0	12	1.4

Table 3.5: 403-9a

The number of fatalities as a result of work-related injury is 0 for more than the 75% of the societies who reported the information. It is important to stress that

this datum is missing for about 20% of the firms. This could be a deliberate lack of reporting to cover such a susceptible information.

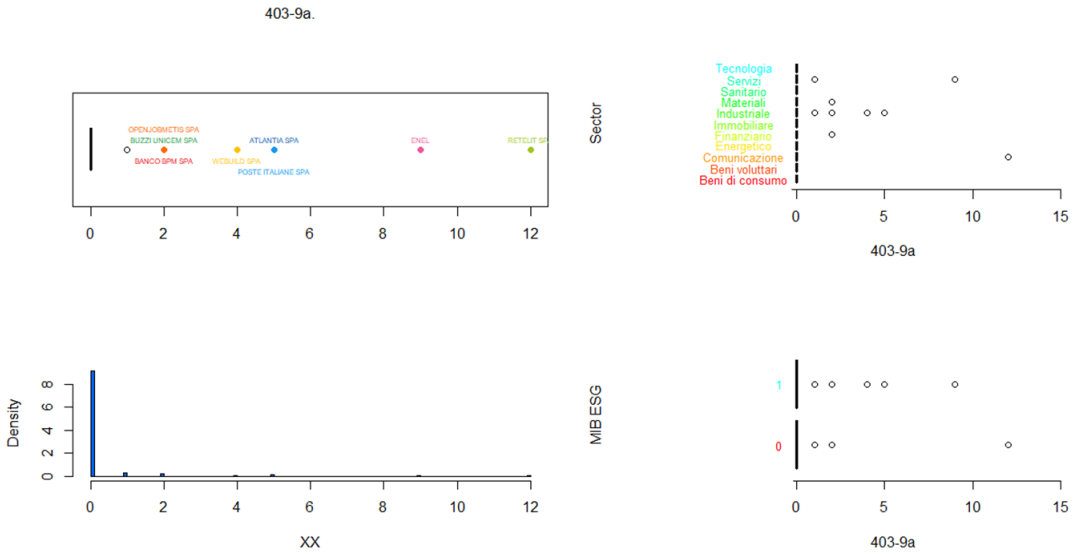


Figure 3.10: 403-9a.

- 403-9b.

	NA (%)	Mean	Min	Max	Std
% Inf	7.3	14.8	0	1145.9	92.8

Table 3.6: 403-9b

The datum result low on average. It is important to underline the fact that, for GRI, un higher rate, is not necessarely a negative factor, instead it can be the result of a better reporting, and can be the sign of a more transparent company.

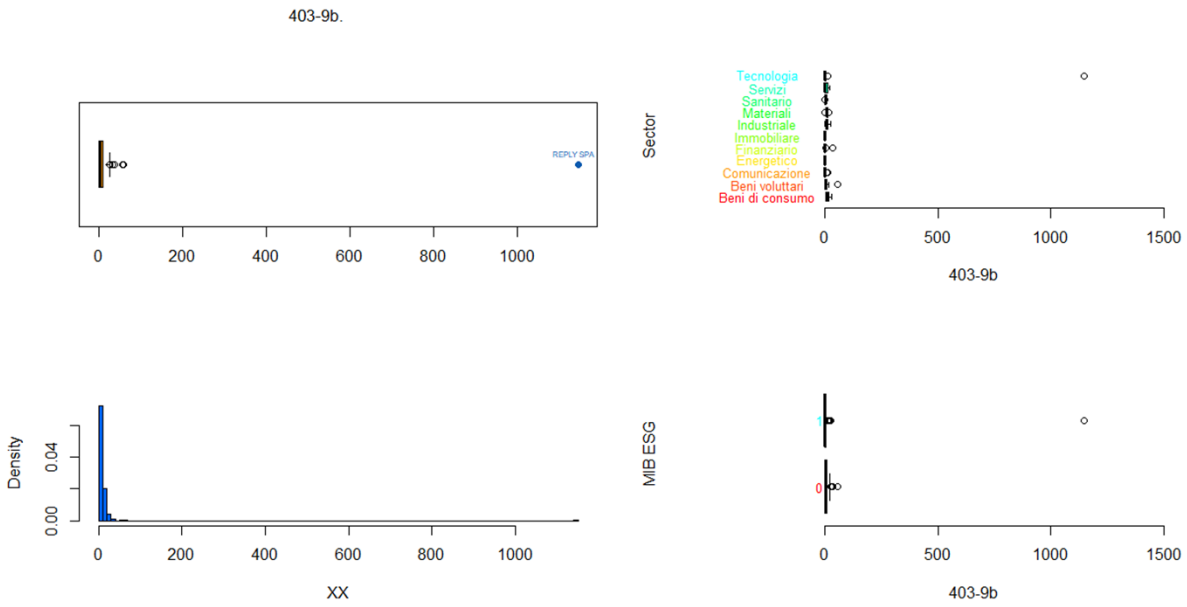


Figure 3.11: 403-9b.

404-1

- 404-1a.

	NA (%)	Mean	Min	Max	Std
D	18.3	16.7	0.5	136.2	17.1
U	18.9	17.9	0.6	121.6	16.3

Table 3.7: 404-1a

The training hours provided seem to be almost equal among men and women. MIB ESG societies and organizations working in the finanziario, sanitario e immobiliare sectors seems to provide more hours of education.

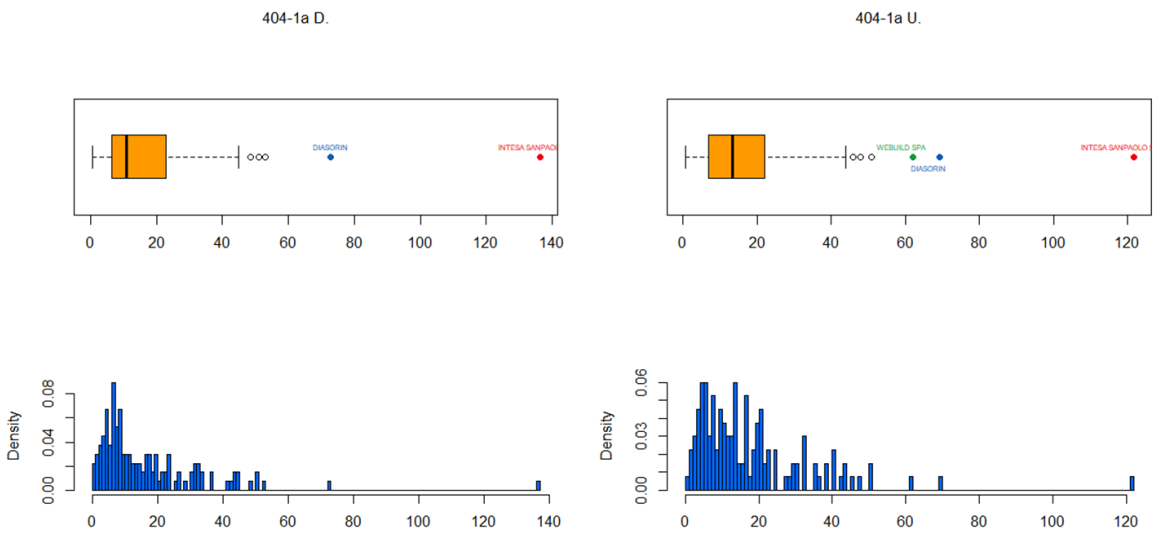


Figure 3.12: 404-1a.

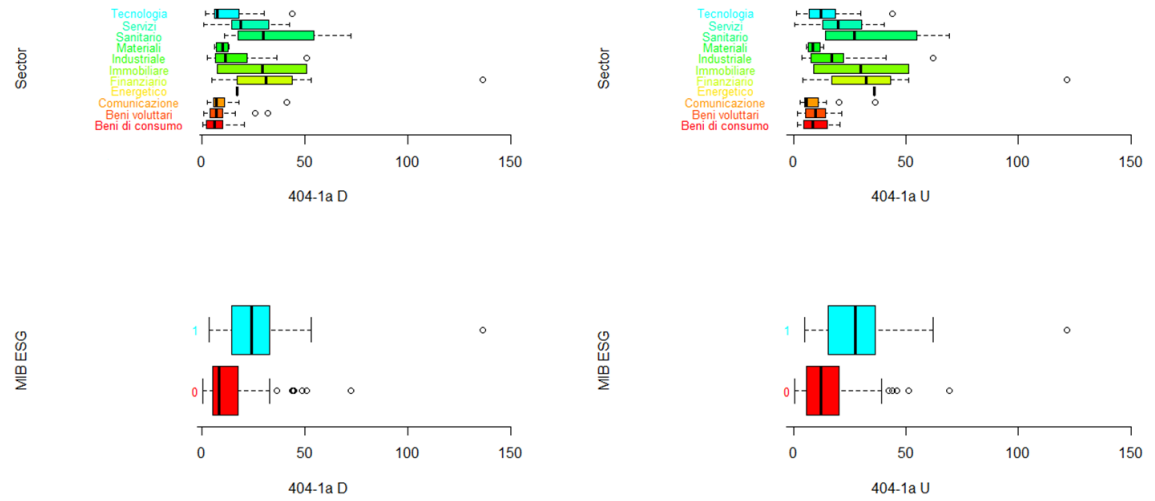


Figure 3.13: 404-1a Sector & MIB ESG.

- 404-1b.

	NA (%)	Mean	Min	Max	Std
Train D	29.3	16.1	0	66	13.9
Train Q	45.7	20.8	0.2	131.2	18.8
Train I	37.2	15.6	0.2	60	12.1
Train O	43.9	10.7	0	66	10.4

Table 3.8: 404-1b

Training hours seems identical for dirigenti, quadri and impiegati,instead they are lower, about 5-10 hours less, for operai.
Dirigenti and quadri are provided more training ours if they work in societies of sector finanziario.While impiegati in sector immobiliare and operai in sector energetico.
MIB ESG organizations,on average, train more their employees in all categories.

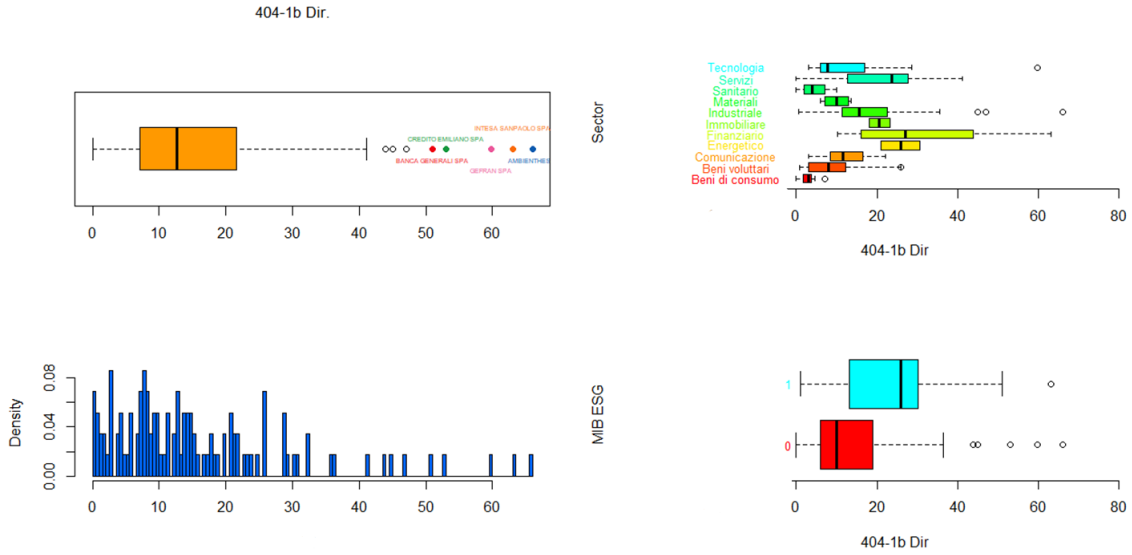


Figure 3.14: 404-1b Dirigenti.

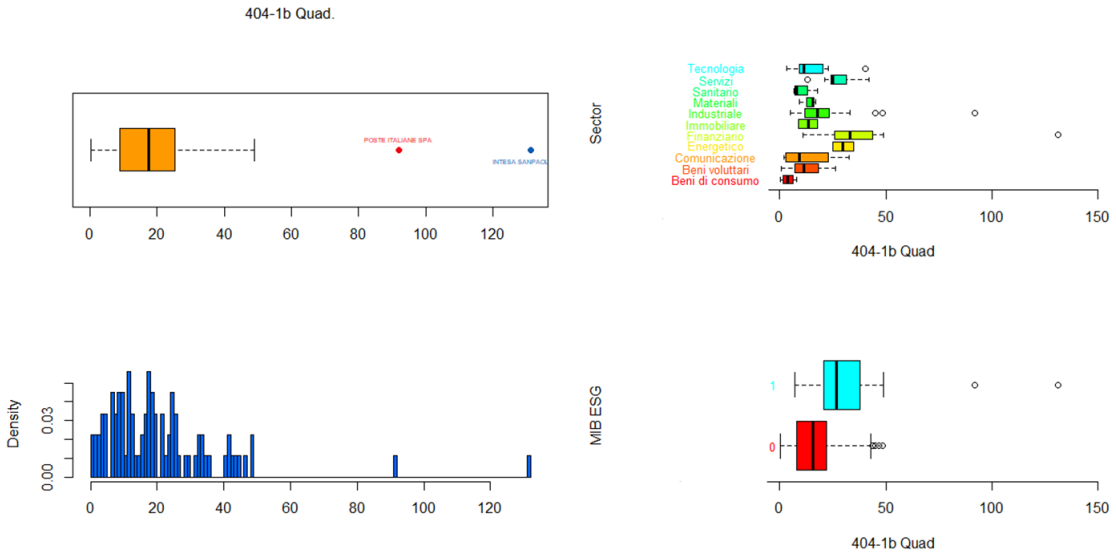


Figure 3.15: 404-1b Quadri.

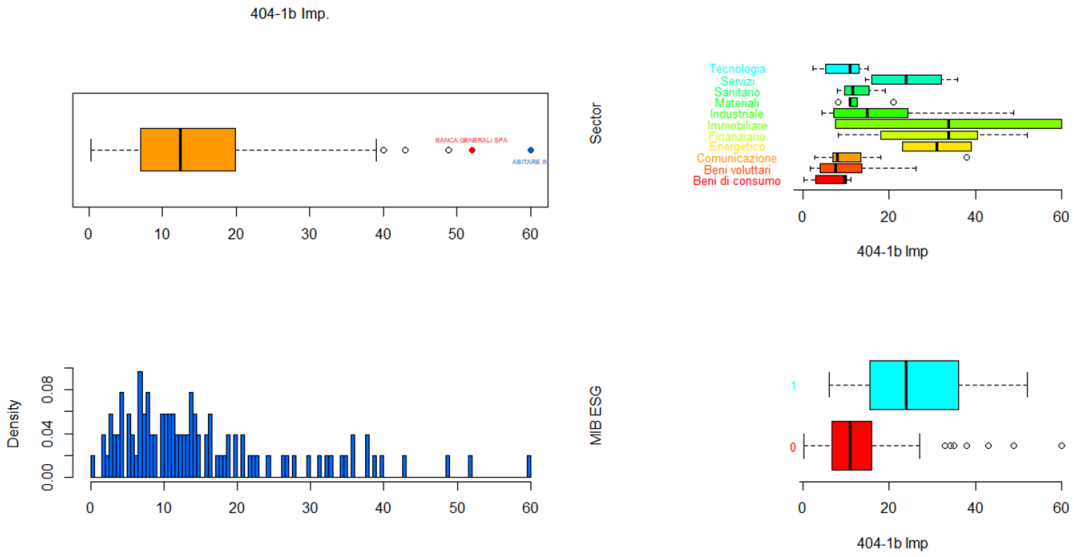


Figure 3.16: 404-1b Impiegati.

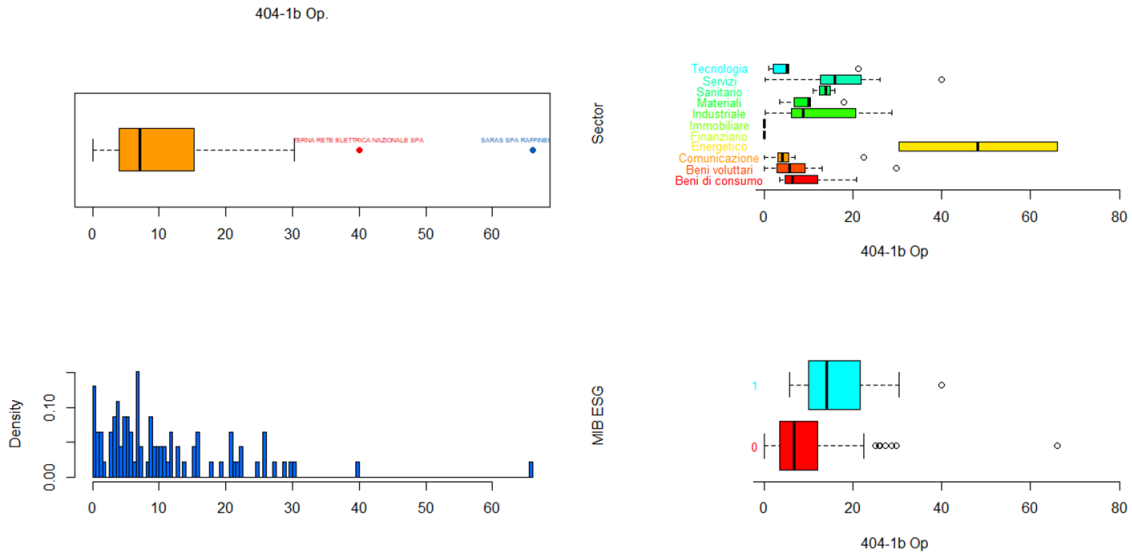


Figure 3.17: 404-1b Operai.

405-1

- 405-1a.

	NA (%)	Mean	Min	Max	Std
% F Gov	2.4	37.7	0	58.3	8.9

Table 3.9: 405-1a

The presence of women in the board of directors is less of 50%, for more than 75% of the firms. This situation does not change neither in a specific sectors nor for the MIB ESG companies.

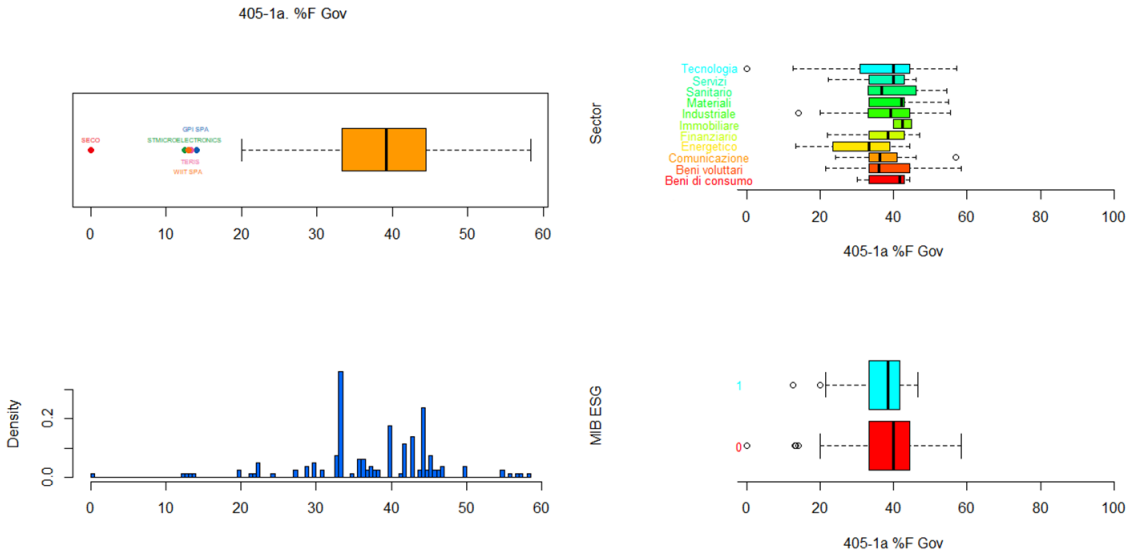


Figure 3.18: 405-1a.

- 405-1b.

	NA (%)	Mean	Min	Max	Std
% G <30	6.1	0.7	0	14.3	2.7
% G 30-50	6.1	24.5	0	75	17.2
% G >50	5.5	72.9	4	100	19.9

Table 3.10: 405-1b

The presence of under-30s on Boards of Directors appears to be almost zero for almost all the companies considered, with a few exceptions of just over 10%. It appears that BoDs are mostly composed of people over 50, the average composition could be around 25 per cent between 30-50 and 75 per cent over 50. It should be noted that MIB ESG companies have a higher average board age.

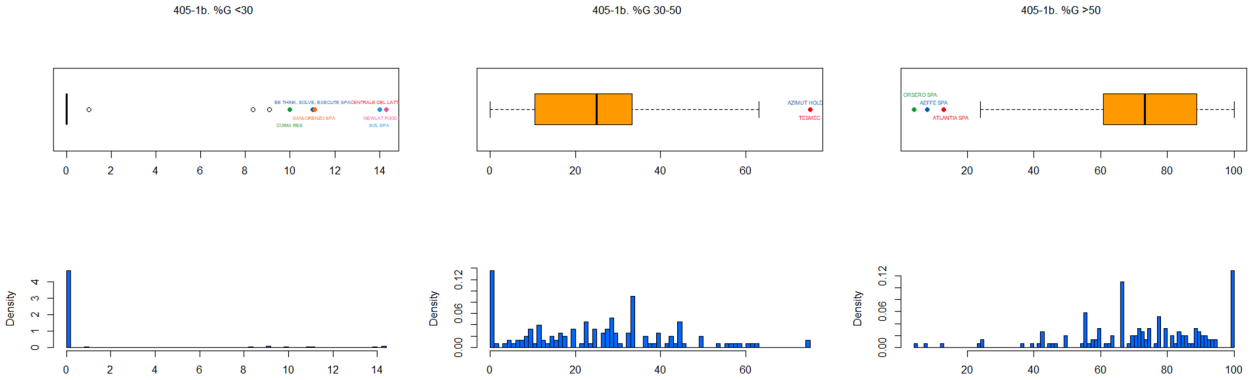


Figure 3.19: 405-1b.

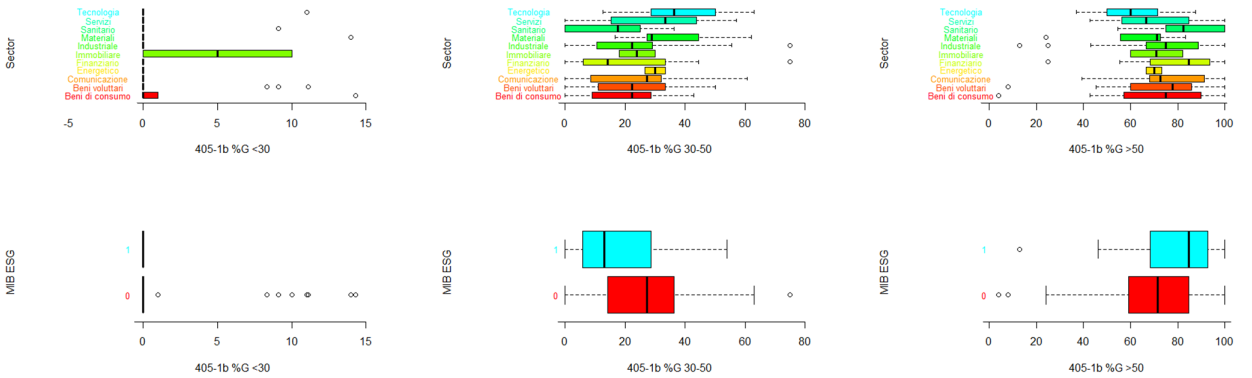


Figure 3.20: 405-1b. Sector & MIB ESG

- 405-1c.

	NA (%)	Mean	Min	Max	Std
% F Tot	1.2	37.1	0.3	79.3	19.1
%F Dir	25	14.6	0	58.9	10.1
%F Qu	42.7	27.4	6.7	59.5	11.9
%F Im	36	42.7	0.4	84	17.9
%F Op	40.2	22.1	0	100	26.1

Table 3.11: 405-1c

The presence of women in the company is below 40% for more than half of the companies surveyed, in some cases even zero. The role with the highest presence

is white-collar workers, where for more than half of the companies, the presence is above 40 %. In the other roles, the average presence is between 20% and 30%.
A further 67 companies presented the figure, but it is difficult to compare them in their entirety.

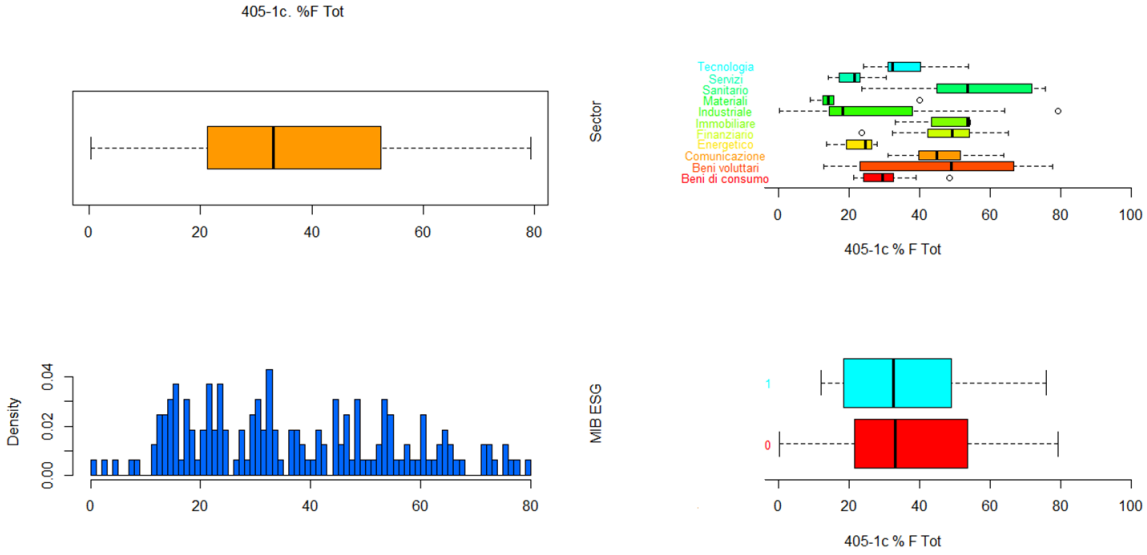


Figure 3.21: 405-1c. % F Tot

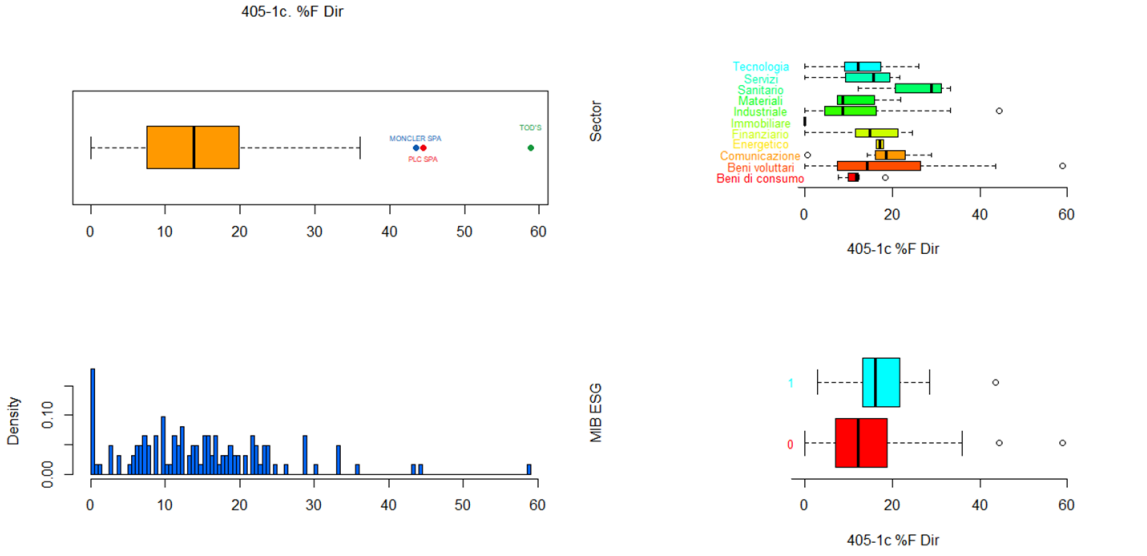


Figure 3.22: 405-1c. % F Dir

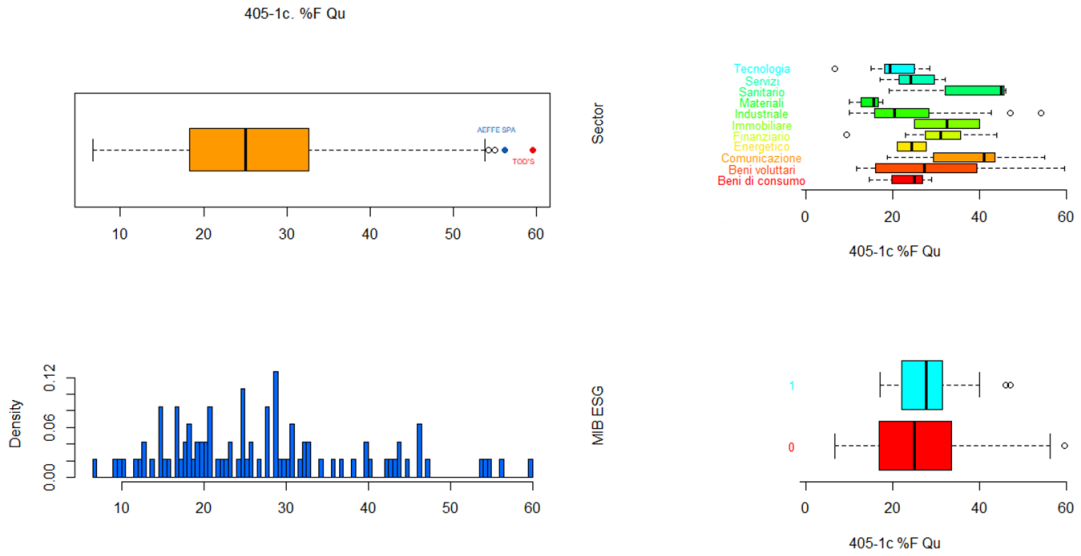


Figure 3.23: 405-1c. % F Qu

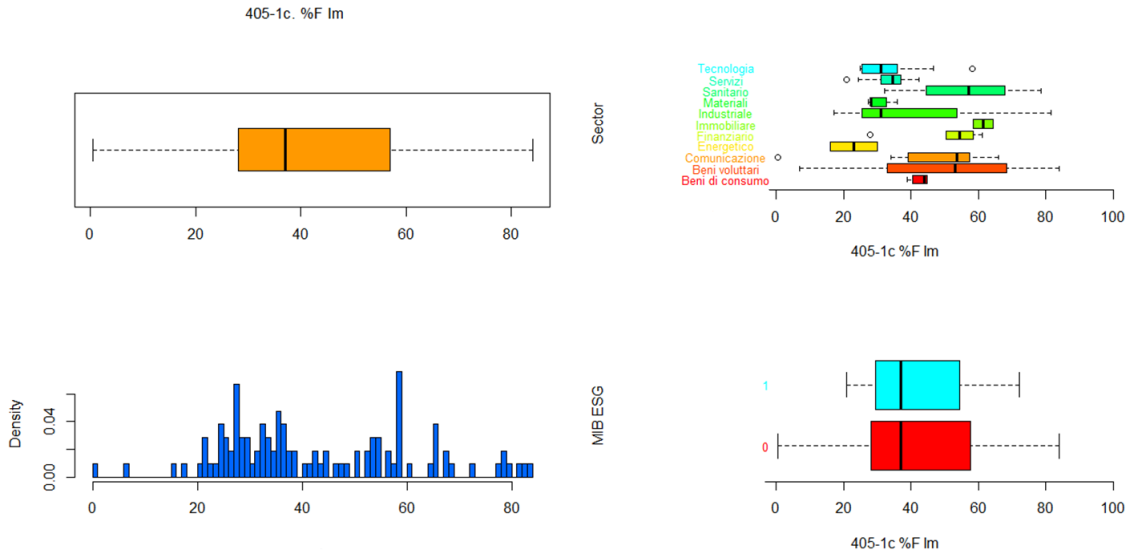


Figure 3.24: 405-1c. % F Im

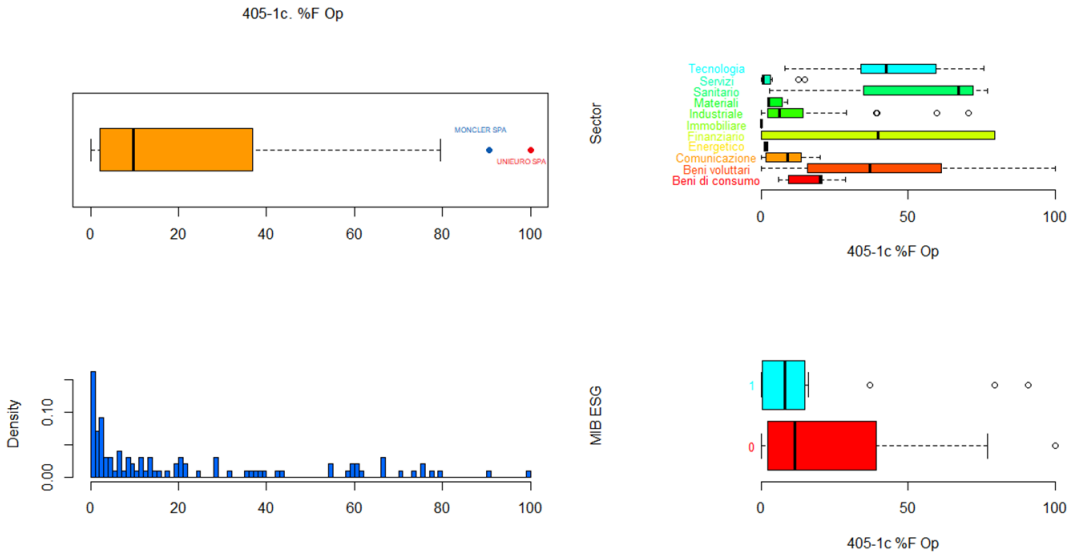


Figure 3.25: 405-1c. % F Op

- 405-1d.
  Dirigenti

	NA (%)	Mean	Min	Max	Std
<30	40.9	0.4	0	10	1.4
30-50	40.9	41.4	0	89.1	16.1
>50	40.9	57.2	1	100	16.9

Table 3.12: 405-1d

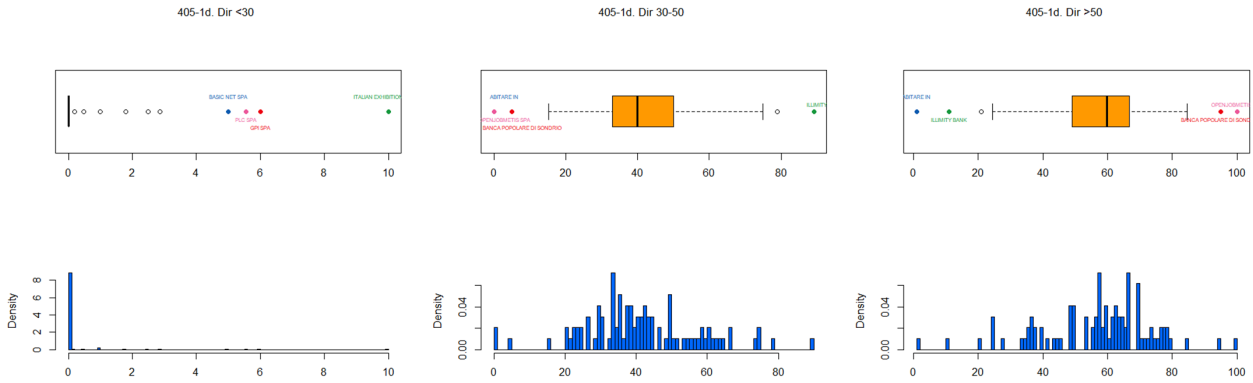


Figure 3.26: 405-1d. Dir

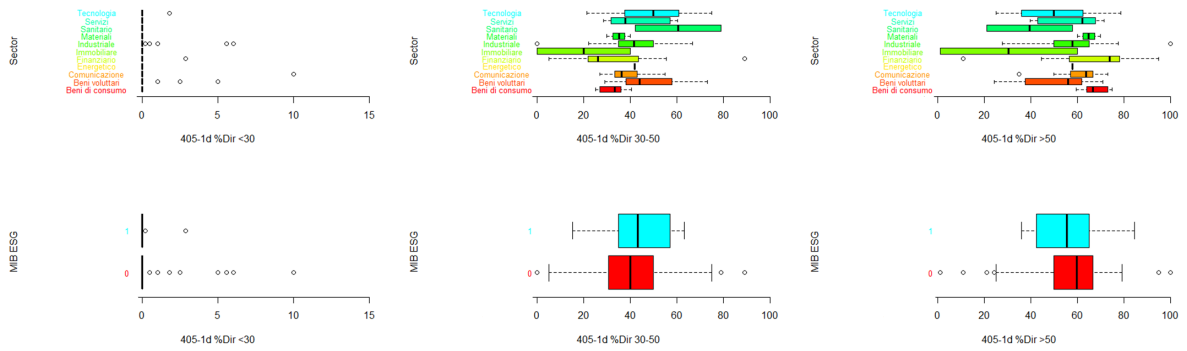


Figure 3.27: 405-1d. Dir Sector & MIB ESG

Quadri

	NA (%)	Mean	Min	Max	Std
<30	53.7	1.7	0	26.9	3.9
30-50	53.7	55.4	1	90	16.4
>50	53.7	41.5	0	81.8	17.3

Table 3.13: 405-1d

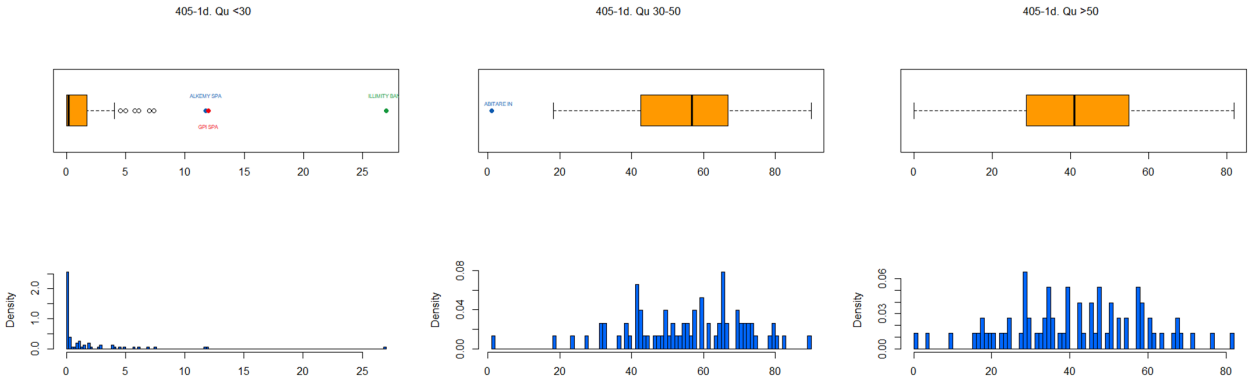


Figure 3.28: 405-1d. Qu

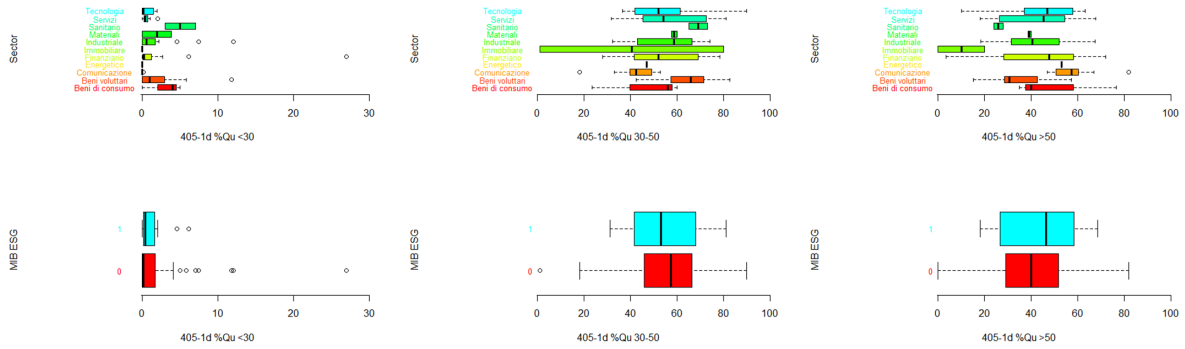


Figure 3.29: 405-1d. Qu Sector & MIB ESG

Impiegati

	NA (%)	Mean	Min	Max	Std
<30	51.2	13.7	0	39	7.7
30-50	51.2	61.7	23.9	84.4	9.7
>50	51.2	24.6	3.9	65.6	12.1

Table 3.14: 405-1d

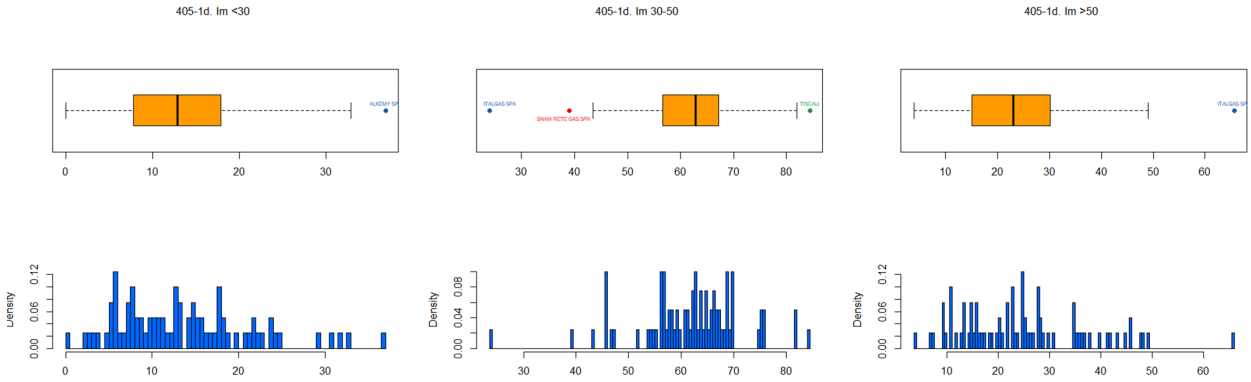


Figure 3.30: 405-1d. Im

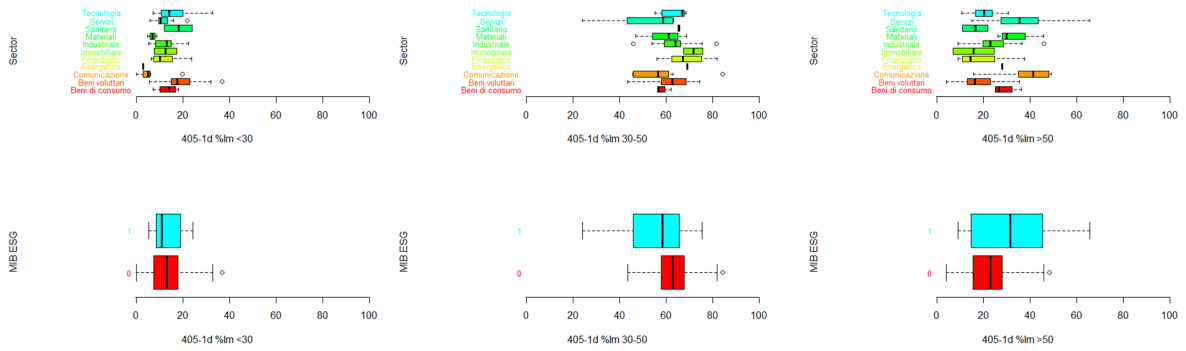


Figure 3.31: 405-1d. Im Sector & MIB ESG

Operai

	NA (%)	Mean	Min	Max	Std
<30	55.5	13.5	0	44.4	8.9
30-50	55.5	54.2	0	100	17.1
>50	55.5	29.7	0	100	15.8

Table 3.15: 405-1d

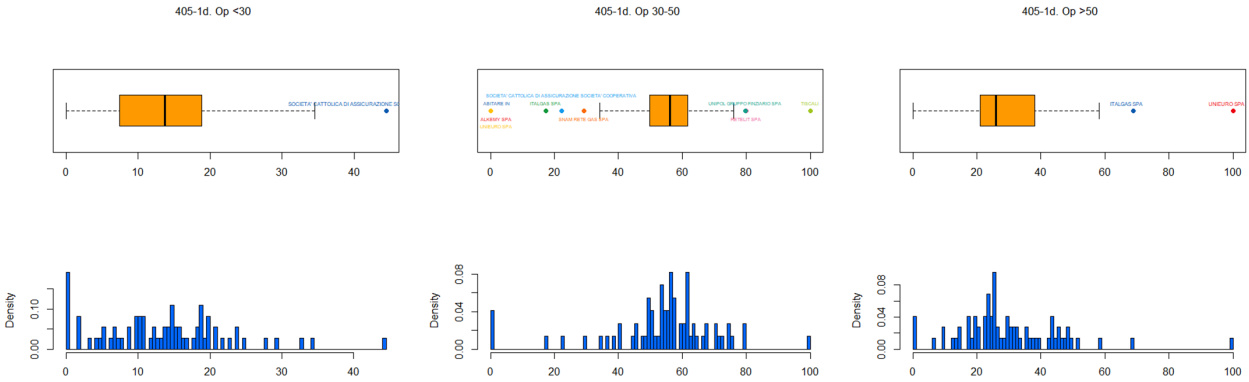


Figure 3.32: 405-1d. Op

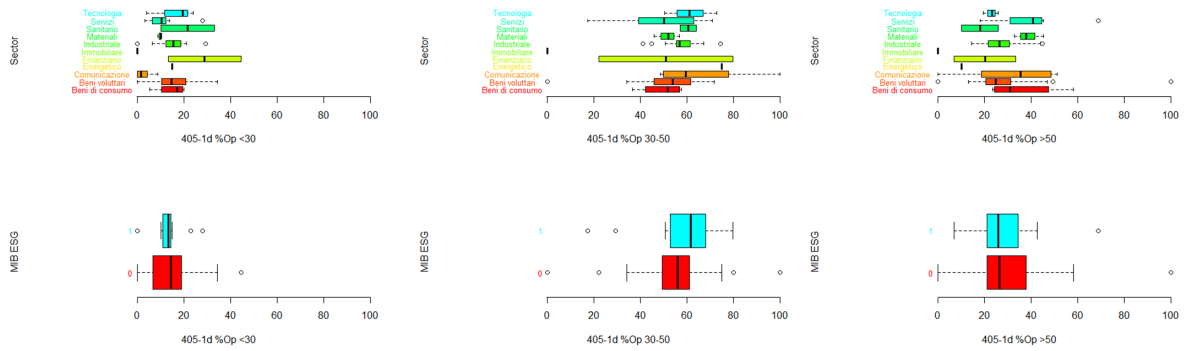


Figure 3.33: 405-1d. Op Sector & MIB ESG

For the higher-level categories, the presence of under-30s is very low or nil. On average, we could say that managers are 40 % between 30-50 and 60 % over 50.
Quadri 3% under 30, 56% between 30-50 e 41% over 50.
In the other categories, the presence of under-30s varies between 10% and 20%, with the 30-50 bracket being the most present.
No specific trends are apparent due to the sector in which a particular company operates or its presence in the MIB ESG index.
There are a further 3 columns in the dataset, which contain data from 47 companies that presented in their dnf data on workers in the various age groups, but in different occupational categories. Again, they are difficult to compare.
A further 3 columns report the percentage of employees by age group. This figure is analysable but less relevant for the purposes of this survey.

405-2

	NA (%)	Mean	Min	Max	Std
Dir	69.5	85.6	39	124.5	15.5
Qu	72.6	93.9	63	112.9	8.8
Im	73.2	91.8	75	122	8.8
Op	80.5	85.9	0	105.2	19.3

Table 3.16: 405-2

For more than 75% of the companies that reported such a figure, the ratio of wages was below the parity threshold for each occupational category.
The variability is lower for companies in the MIB ESG, but the gap persists, despite

appearing slightly smaller.
It should be emphasised that this is the least presented data in absolutes, an absence of more than 50%, and presented in a few cases in a manner different from that desired.
An additional column was introduced showing the data of 17 companies that reported the ratio of wages for their employee categories. For these, the same considerations apply as for the compliant figure, in fact, on average, it is below the equality.

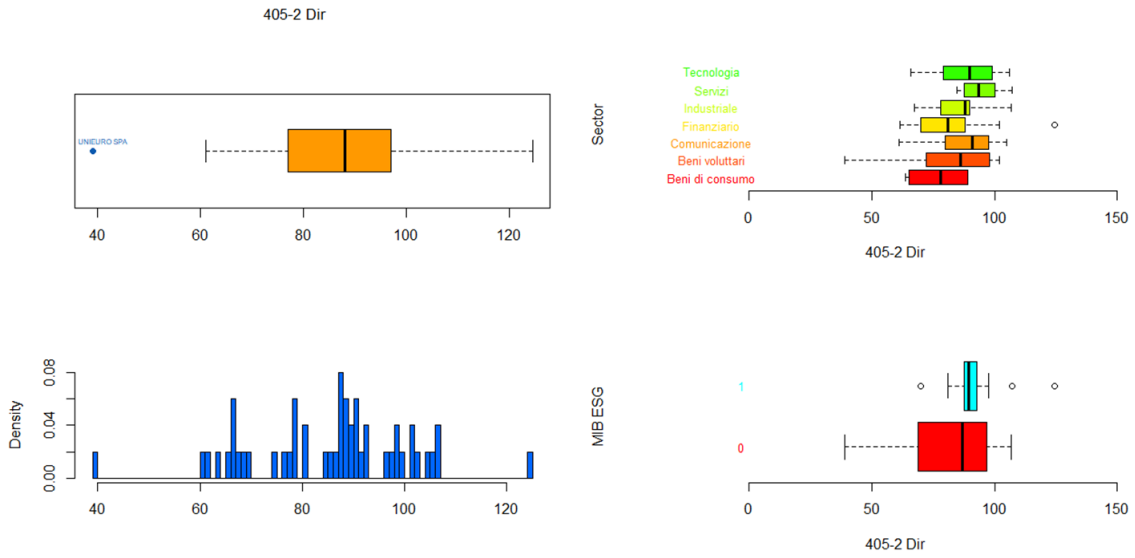


Figure 3.34: 405-2. Dir

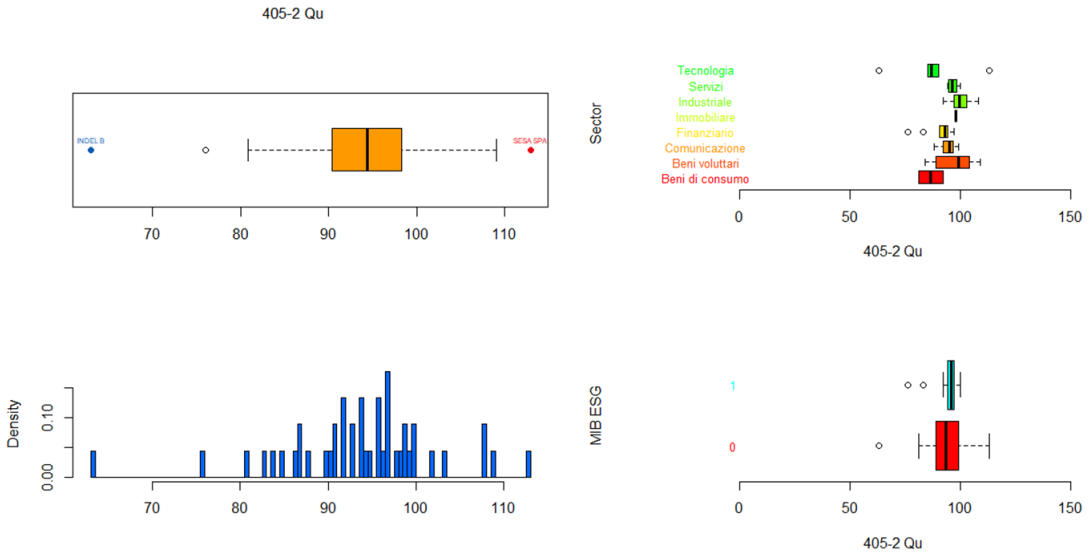


Figure 3.35: 405-2. Qu

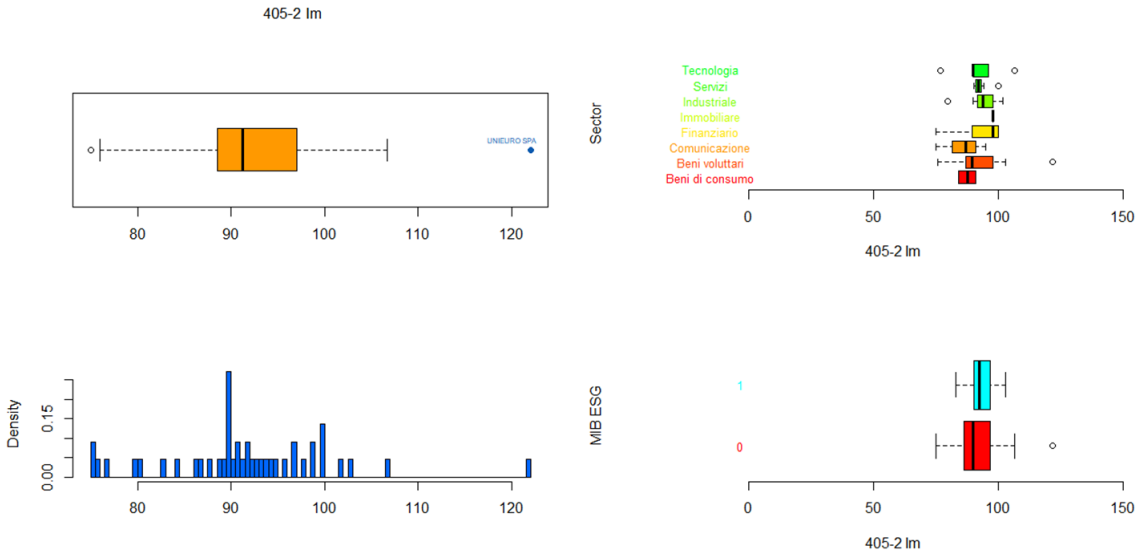


Figure 3.36: 405-2. Im

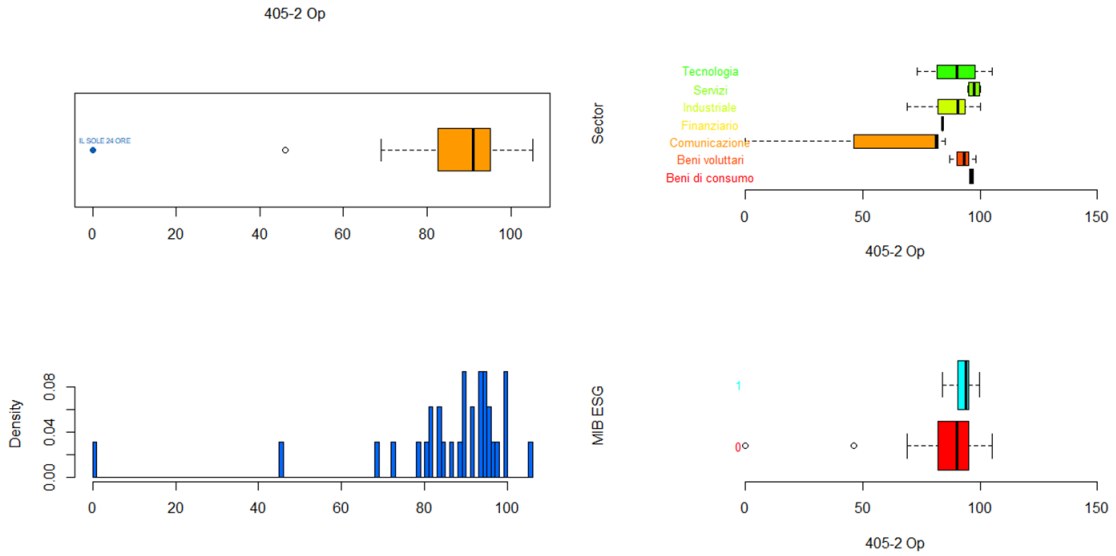


Figure 3.37: 405-2. Op

Additional remarks

The male-female disparity appears more pronounced especially in more decision-making roles such as board of directors and management.

The presence of under-30s appears to be low in most companies, however, given the recruitment in 2020, there seems to be an intention to increase the share of this group.

The MIB ESG companies seem to be the most virtuous overall, but this does not mean that they are adhering to good social standards regarding employee welfare.

4 | Network analysis

In the previous chapter, initial considerations were made on the data collected, highlighting poor reporting by companies.
Trends emerged, which, on average, describe the social pillar behaviour of Italian companies.
The final step is to try to find out which companies are most caring about social issues.
For this research, we will use the network theory tools discussed above.

4.1. Data Prepocessing

n order to build the network, the first step is to deal with the large number of missing values.
Before proceeding with data cleaning, let us consider the following working hypothesis:

Proposition 4.1. *A bad reporting is signal of a bad effort in the integration of Social factors.*

Thus, by excluding companies with a high number of discrepancies and/or reporting failures, we obtain the set of companies that can be considered poorly performing.
We then exclude those GRI indices for which there are few samples.
This reduces the analysis to only 91 companies, out of the initial 164. These constitute the nodes of the network and are as follows:

1. A2A SPA
2. ACEA SPA
3. AEROPORTO GUGLIELMO MARCONI DI BOLOG SPA
4. ASCOPIAVE SPA
5. ATLANTIA SPA
6. AVIO SPA
7. AZIMUT HOLDING
8. BANCA CARIGE
9. BANCA FARMAFACTORING SPA
10. BANCA GENERALI SPA
11. BANCA IFIS SPA
12. BANCA MEDIOLANUM SPA

13. BANCA MONTE DEI PASCHI DI SIE SPA
14. BANCA POPOLARE DELL'EMILIA ROMAG, SOCIETA' COOPERATIVA
15. BANCA POPOLARE DI SONDRIO
16. BE THINK, SOLVE, EXECUTE SPA
17. BEGHELLI
18. BIESSE SPA
19. BREMBO
20. CAIRO COMMUNICATION SPA
21. CALTAGIRONE SPA
22. CAMPARI
23. CAREL INDUSTRIES SPA
24. CEMBRE SPA
25. CENTRALE DEL LATTE D'ITALIA
26. CERVED GROUP SPA
27. CNH INDUSTRIAL
28. CSP INTERTIOL
29. DATALOGIC
30. DOVALUE
31. ELICA
32. EMAK
33. ENEL
34. EV
35. FALCK RENEWABLES
36. FERRARI
37. FIERA MILANO
38. FINECOBANK
39. FNM SPA
40. GPI SPA
41. IGD - SIIQ
42. IL SOLE 24 ORE
43. ILLIMITY BANK
44. INDEL B
45. INWIT
46. ITALGAS SPA
47. ITALIAN EXHIBITION GROUP
48. ITALMOBILIARE
49. IVS GROUP
50. LA DORIA SPA
51. LANDI RENZO SPA
52. LEONARDO SPA
53. LUVE SPA
54. MARR SPA
55. MEDIASET SPA
56. MEDIOBANCA - BANCA DI CREDITO FINZIARIO SPA
57. MONCLER SPA
58. MONDADORI EDITORE SPA
59. MONRIF SPA
60. NEODECORTECH SPA
61. NEWLAT FOOD SPA
62. ORSERO SPA

63. PIOVAN SPA
64. PIQUADRO SPA
65. PIRELLI SPA
66. PLC SPA
67. RAI WAY SPA
68. RATTI SPA
69. RCS MEDIAGROUP SPA
70. RECORDATI INDUSTRIA CHIMICA E FARMACEUTICA SPA
71. RENO DE MEDICI SPA
72. RETELIT SPA
73. SABAF SPA
74. SAFILO GROUP SPA
75. SALCEF GROUP SPA
76. SNAM RETE GAS SPA
77. SOCIETA' CATTOLICA DI ASSICURAZIONE SOCIETA' COOPERATIVA
78. SOL SPA
79. STELLANTIS
80. TAS
81. TELECOM ITALIA
82. TERNA RETE ELETTRICA NAZIONALE SPA
83. TESMEC
84. TINEXTA
85. TISCALI
86. TOD'S
87. TXT E-SOLUTIONS SPA
88. UNICREDIT SPA
89. UNIEURO SPA
90. UNIPOL GRUPPO FINZIARIO SPA
91. UNIPOLSAI SPA

The numerical index associated with them will be used in the network representations for identification purposes.
It is worth highlighting the fact that of the companies considered, 22 (out of 28) belong to the MIB ESG index.

4.2. Matrix of weights

The aim is to build a weighted network, and the weighted links will be the result of some similarity among the firms. We will consider a similarity based on distance as defined in chapter 1.
The first consideration concerns the scale of the data.. If we consider a distance based on the magnitude scale, for instance Euclidean or Manhattan, the following issues verify.
Companies we are analyzing differ in terms of capitalization and size: STELLANTIS has

over 400,000 employees, while ASCOPIAVE SPA has just over 400, they seem to be hardly comparable in some entries.
More in detail we can highlight the global situation.
Consider the dendograms in Figure 4.1 and in Figure 4.2, obatined using Euclidean distance and Manhattan distance respectively, and using Complete, Average and Single linkage to update the distance matrices.

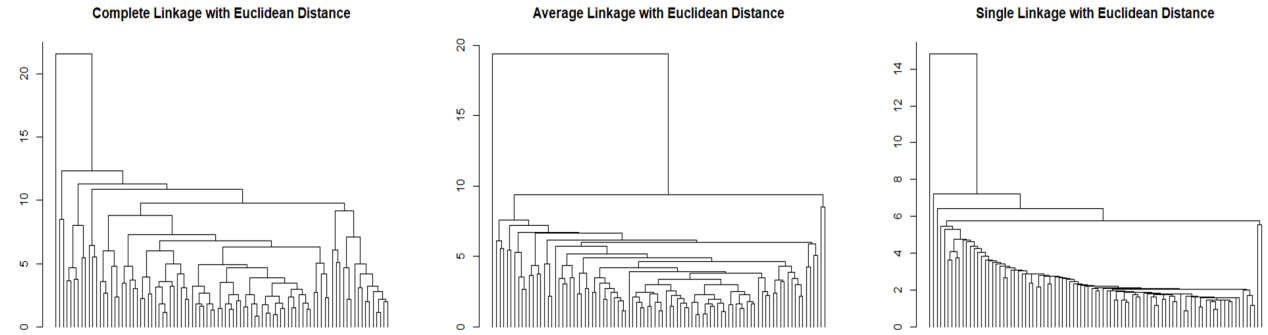


Figure 4.1: Euclidean dendograms

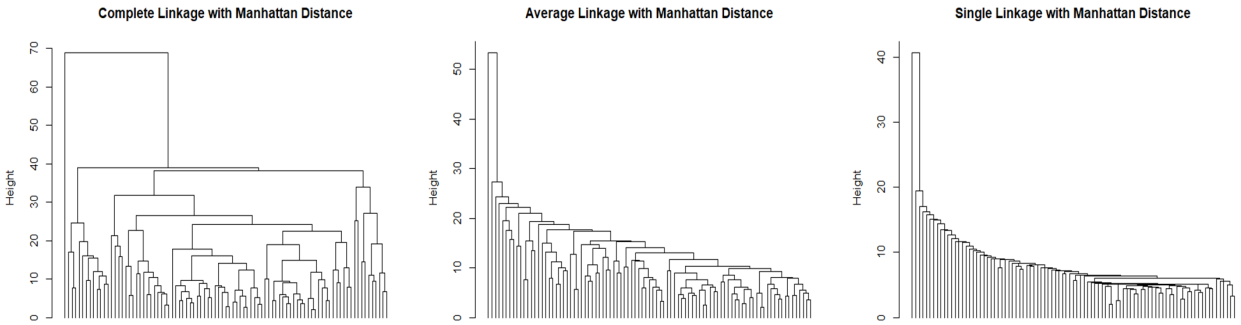


Figure 4.2: Manhattan dendograms

In both scenarios, the differences among the firms seem obvious, in fact, they organize in a large number of clusters, each one containing few(in some cases just one) companies.
Even standardizing the entries the situation do not improve.
The most reasonable approach seems to use a distance based on the shape, i.e the Pearson correlation distance.
We recall the definition and introduce the Pearson correlation similarity:

Definition 4.1. ***Pearson correlation distance and similarity***

$$d(X,Y) = 1 - r(X,Y)$$

where,

$$r(X,Y) = \frac{\sum_{i=1}^{p}(x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{p}(x_i - \overline{X})^2 \sum_{i=1}^{p}(y_i - \overline{Y})^2}}$$

The associated similarity is

$$s_d(X,Y) = \frac{1}{1 + d(X,Y)} = \frac{1}{2 - r(X,Y))}$$

It is possible to see how much the different firms are correlated in Figure 4.3.

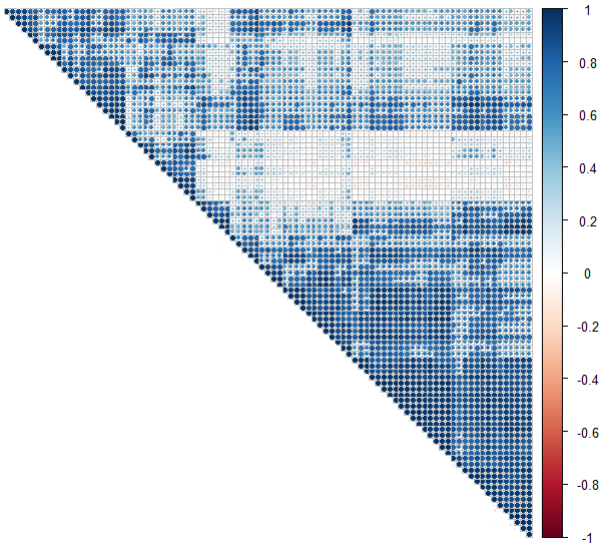


Figure 4.3: Corrplot: correlation among the 91 companies

We compute the Pearson distance and as before we want to see if it seems a reasonable choice in order to build the network.
As we can see in Figure 4.4 the situation is more clear, and a first look at the dendograms seems to display two different clusters.
Further details will be reported in the following section.

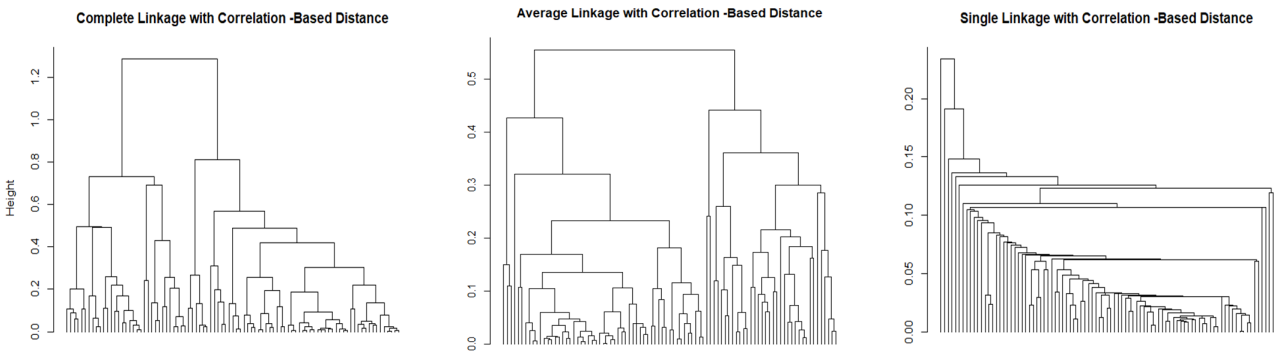


Figure 4.4: Pearson correlation dendograms

Chosen the similarity, we build the network $\mathcal{G}_w = (\mathcal{N}, \mathcal{L}, \mathcal{W})$.

4.3. Network analysis

The weights matrix define the links:

$$W = (w_{ij}) = \frac{1}{2 - r_{ij}} > 0$$

Since weights are strictly positive, all nodes in $\mathcal{N}$ are connected.
The network was built using the *R package igraph*[9].
In Figure 4.5 is represented the obtained network.
For graphical purposes, only links with a higher than average weight are shown.
The size of the nodes is proportional to their strength.

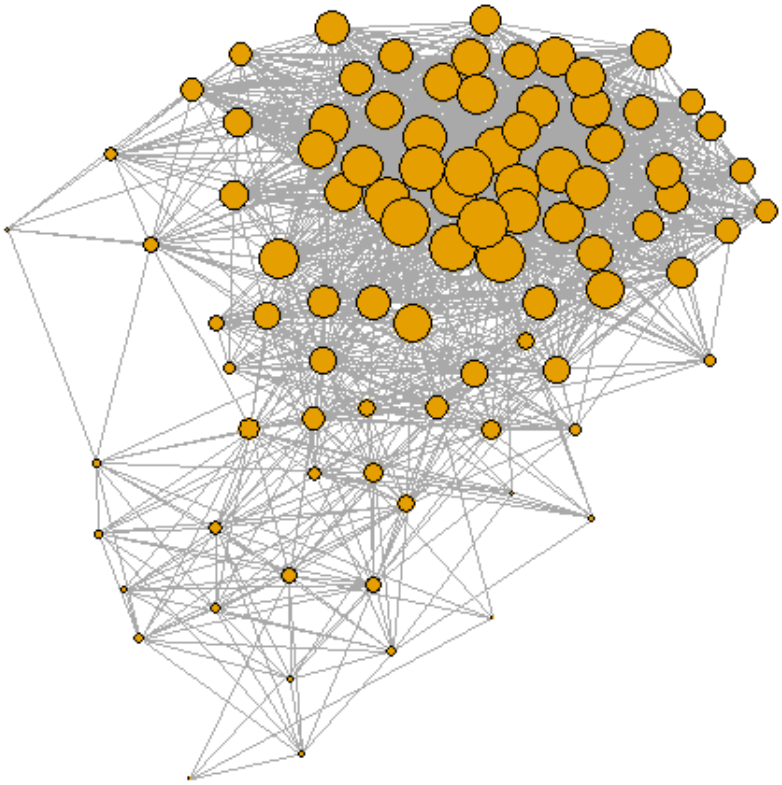


Figure 4.5: Network

4.4. Candidate communities

The real goal is to find sub-networks of companies that operate in a similar way, so that we can then understand how well a company complies with the social pillar depending on how it is clustered.
The tool used in this analysis is the community detection.
The procedure adopted is outlined below:

- Find a candidate community with different methodologies.

- Asses the goodness of a partition,i.e, verify if it is an α-partition, for $\alpha = 0.5$ fixed a priori.

Two possible partition are taken into consideration.

1. Hierarchical clustering

In the selection of the similarity measure, the tool used to visualize whether the selected measure was meaningful or not was the dendogram.

This trivially lead us to consider the clusters which can be obtained via hierarchical clustering.

Consider the case we adopt complete linkage to update the distance matrix.

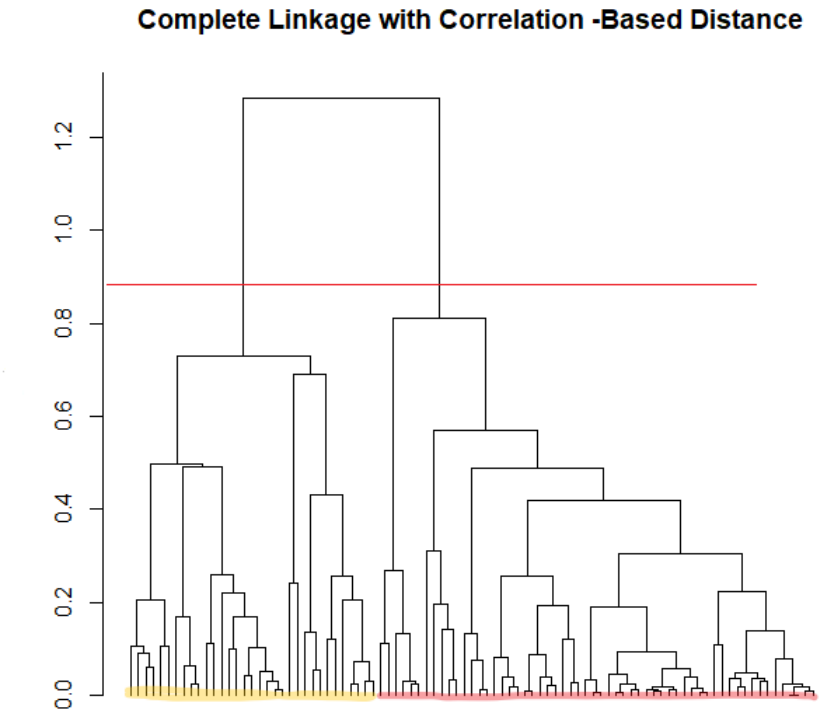


Figure 4.6: Dendogram cut at an height about 0.9 produce two different clusters

We obtain $\mathcal{C}_1$ which is composed by 58 societies, and $\mathcal{C}_2$ by 33.

The representation on the network is the following:

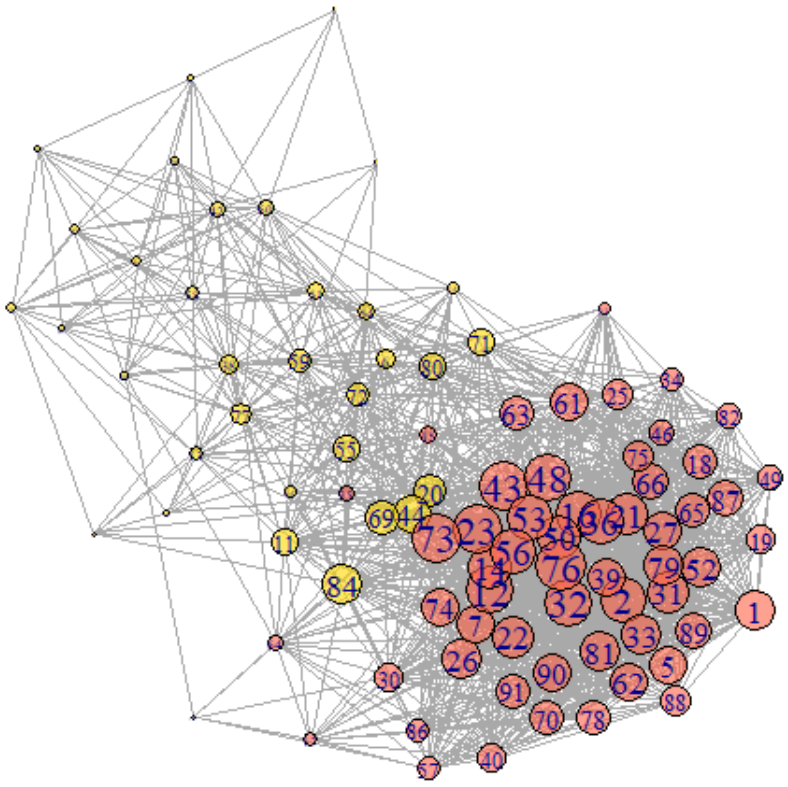


Figure 4.7: $\mathcal{C}_1$ is the red partition; $\mathcal{C}_2$ is the golden one.

To asses the quality we can evaluate the persistence probabilities.
Reminding that in the case of a weighted undirected network they can be easily computed:

$$u_{cc} = \frac{\sum_{i,j \in \mathcal{C}_c} w_{ij}}{\sum_{i \in \mathcal{C}_c} s_i}$$

$$u_{11} \approx 0.65 \qquad u_{22} \approx 0.35$$

It follow that $\mathcal{C}_1$ is an α-partition and $\mathcal{C}_2$ it is not.
The comunities do not have interesting peculiarities since composed by companies belong-

ing to different sectors.

By the way 19 companies out of the 22 in the MIB ESG index are member of $\mathcal{C}_1$.

2. Modularity optimization

Another useful technique in finding communities is relying on the modularity optimization. Thank to the *igraph package* it is easy to find this optimal partition. We obtain $\mathcal{C}_1$ composed by 52 nodes and $\mathcal{C}_2$ by 39.

If we compare with the previous case the obtained result are different just by comparing the sizes. But the difference is minimal since just 6 companies are classified in different cluster in the two scenarios.

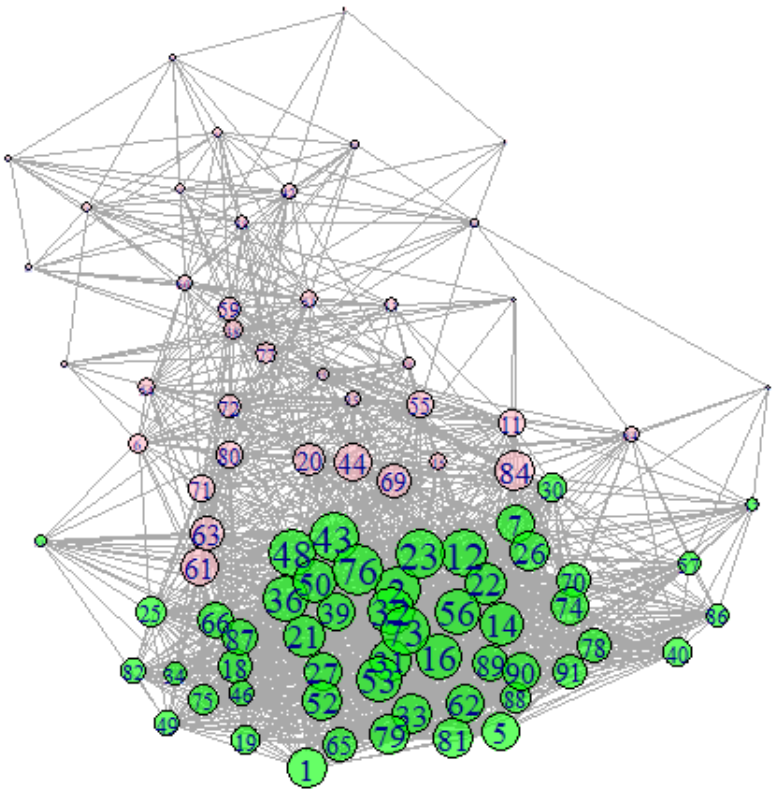


Figure 4.8: $\mathcal{C}_1$ is the green partition; $\mathcal{C}_2$ is the pink one.

The persistence probabilities of the clusters are:

$$u_{11} \approx 0.58 \qquad u_{22} \approx 0.42$$

The difference in the persistence probabilities seems material, since a little migration of nodes make both valus closer to the selected α. By the way $\mathcal{C}_2$ is still not an α-partition. For what concern the societies in the MIB ESG index, only a single company moves to another community, in fact we have 18 firms in $\mathcal{C}_1$ and 4 in $\mathcal{C}_2$.

4.5. Results

The network and the process to obtain it allows us to divide the companies considered into three macro-categories:

- 73 companies reporting numerous deviations from standards, therefore considered poorly performing.
- The two communities obtained with the algorithms considered.

The most significant result is that the companies in the MIB ESG index are mostly classified in the same cluster.
Comparing with the statistics analysed in the previous chapter, it appears that companies in the MIB ESG cluster are on average more virtuous.
This would seem to imply that companies classified in the same cluster are more compliant with good social policies.
Let us try to find further confirmation by looking at the average GRI values according to community membership.

KPIs	Mean $\mathcal{C}_1^1$	Mean $\mathcal{C}_2^1$	Mean $\mathcal{C}_1^2$	Mean $\mathcal{C}_2^2$
N Ass 30	573.5	25.3	632.8	30.6
N Ass 30-50	542.9	35.7	598.7	39.3
N Ass 50	86.9	7.4	95.3	8.5
N Ass F	417.8	26.8	458.6	32.5
N Ass U	785.0	41.2	867.6	45.5
N Turn 30	4.5	2.6	4.3	3.1
N Turn 30-50	6.7	4.1	6.0	5.3
N Turn 50	3.9	3.7	3.8	3.9
N Turn F	5.5	3.7	4.9	4.9
N Turn U	9.0	6.1	8.5	7.1
N Inf	0.26	0.4	0.29	0.33
% Inf	6.8	5.9	6.8	6.1
Train F	15.0	14.9	15.0	14.8
Train U	16.5	15.8	16.5	15.9
Train D	15.1	17.9	15.4	17.2
% F Gov	37.3	40.4	37.3	39.9
% G 30	1.1	0	0.95	0.37
% G 30-50	26.1	24.8	26.3	24.6
% G 50	69.6	75.2	69.1	75.0
% F Tot	33.6	40.5	32.9	40.3

Table 4.1: Mean of the obtained communities. Superscript 1: Hierarchical clustering; Superscript 2: Modularity optimization.

- Data on recruitment appear to be misleading when considered in absolute terms. However, $\mathcal{C}^1$ companies seem to hire more under-30s while $\mathcal{C}^2$ companies 30-50.
- The average value of turnover records is higher for $\mathcal{C}^1$ companies. As already discussed these contain a double meaning: workers dissatisfaction or change in the corporation structure.
- The number of accidents is lower for $\mathcal{C}^1$ while the rate is higher. This indicates better reporting of this significant datum.
- The average training hours by gender are higher in $\mathcal{C}^1$. As already analysed for training, the tendency is to train more in the most prestigious categories. In this case, looking at the data by gender, therefore, $\mathcal{C}^1$ companies seem to train their entire staff for more hours.

- The presence of women is considerably higher in $\mathcal{C}^2$ companies, both when considering the composition of the board of directors and the total presence in the company.
- Companies in the $\mathcal{C}^1$ group have on average a lower age of the board of directors, in some cases even under 30 in these positions.

Thus, a completely favourable situation for $\mathcal{C}^1$ companies does not emerge.
In general, it seems that $\mathcal{C}^1$ companies offer more opportunities to young people, in rare cases even in leading positions, provide more training for their employees.
To the credit of the $\mathcal{C}^2$ companies is the fact that they guarantee more opportunities for women.
It must be remembered, however, that the situation remains rather far from gender equality.
We could conclude that on average, companies in $\mathcal{C}_1$ and thus more similar to companies in the MIB ESG index are more socially oriented.
However, major doubts emerge on the MIB ESG, as 6 companies are considered poorly performing, and at least 3 more are clustered in the less performing cluster.
Overall, there are 9-10 companies in the MIB ESG, i.e more than 32%, that do not rank among the most socially responsible.
The composition of this index is therefore not entirely clear.
Obviously, considerations of the Environmental and Governance factors are necessary for an overall assessment.

Bibliography

[1] Soggetti che hanno pubblicato la dichiarazione non finanziaria. https://www.consob.it/web/area-pubblica/soggetti-che-hanno-pubblicato-la-dnf#_ftnref2.

[2] https://www.esgcorporatedata.com/osservatorio/.

[3] Consolidated set of the gri standards 2021. https://www.globalreporting.org/how-to-use-the-gri-standards/gri-standards-english-language/.

[4] Mib esg index composition. https://www.consob.it/web/area-pubblica/soggetti-che-hanno-pubblicato-la-dnf#_ftnref2.

[5] Gri 401 occupazione 2016. https://www.globalreporting.org/standards/media/2139/italian-gri-401-employment-2016.pdf, .

[6] Gri 403 salute e sicurezza sul lavoro 2018. https://www.globalreporting.org/standards/media/2142/italian-gri-403-occupational-health-and-safety-2018.pdf, .

[7] Gri 404 formazione e istruzione 2016. https://www.globalreporting.org/standards/media/2143/italian-gri-404-training-and-education-2016.pdf, .

[8] Gri 405 diveristà e pari opportunità 2016. https://www.globalreporting.org/standards/media/2144/italian-gri-405-diversity-and-equal-opportunity-2016.pdf, .

[9] igraph r package. https://igraph.org/r/.

[10] S. E. Ahnert, D. Garlaschelli, T. M. Fink, and G. Caldarelli. Ensemble approach to the analysis of weighted networks. *Physical Review E*, 76(1):016101, 2007.

[11] A. Barabasi. *Network Science*. Cambridge University Press, 2016. URL http://networksciencebook.com/.

[12] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[13] P. Berkhin. *A Survey of Clustering Data Mining Techniques*, pages 25–71. Springer Berlin Heidelberg, 2006.

[14] M. Billio, M. Costola, I. Hristova, C. Latino, and L. Pelizzon. Inside the esg ratings:(dis) agreement and performance. *Corporate Social Responsibility and Environmental Management*, 28(5):1426–1445, 2021.

[15] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics reports*, 424(4-5):175–308, 2006.

[16] S. Dulli, S. Furini, and E. Peron. *Data mining: metodi e strategie*. Springer Science & Business Media, 2009.

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

[18] S. Fortunato and D. Hric. Community detection in networks: A user guide. *Physics reports*, 659:1–44, 2016.

[19] C. Fu and J. Yang. Granular classification for imbalanced datasets: A minkowski distance-based method. *Algorithms*, 14(2):54, 2021.

[20] K. A. S. Immink and J. H. Weber. Minimum pearson distance detection for multilevel channels with gain and/or offset mismatch. *IEEE Transactions on Information Theory*, 60(10):5966–5974, 2014.

[21] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[22] R. A. Johnson, D. W. Wichern, et al. *Applied multivariate statistical analysis*, volume 6. Pearson London, UK:, 2014.

[23] L. Lü and T. Zhou. Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications*, 390(6):1150–1170, 2011.

[24] P. Matos. Esg and responsible institutional investing around the world: A critical review. 2020.

[25] M. E. Newman. The structure and function of complex networks. *SIAM review*, 45 (2):167–256, 2003.

[26] M. E. Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[27] S. Ontañón. An overview of distance and similarity functions for structured data. *Artificial Intelligence Review*, 53(7):5309–5351, 2020.

[28] C. Piccardi. Finding and testing network communities by lumped markov chains. *PloS one*, 6(11):e27028, 2011.

[29] K. Steinhaeuser and N. V. Chawla. Identifying and evaluating community structure in complex networks. *Pattern Recognition Letters*, 31(5):413–421, 2010.

[30] M. Tantardini, F. Ieva, L. Tajoli, and C. Piccardi. Comparing methods for comparing networks. *Scientific reports*, 9(1):1–19, 2019.

[31] X. F. Wang and G. Chen. Complex networks: small-world, scale-free and beyond. *IEEE circuits and systems magazine*, 3(1):6–20, 2003.

List of Figures

1.1 Minkowski distances in a two dimensional space . . . 3
1.2 Intra and inter cluster distances . . . 8
1.3 dataset . . . 11
1.4 Dendogram, output of the hierarchical clustering algorithm on the toy data set. . . . 11
1.5 First four steps of the hierarchical clustering algorithm on the toy data set. 12
1.6 Clusters obtained cutting the dendogram at two different heights. . . . 13
1.7 Dendograms obtained considering the four linkage presented. . . . 14

2.1 Weighted undirected graph . . . 20
2.2 Poisson degree distribution . . . 23
2.3 (**A**) The collaboration graph of movie actors. $\gamma \approx 2,3$. (**B**) WWW. $\gamma \approx 2,1$. (**C**) Electrical power grid of western US. $\gamma \approx 4$. . . . 24

3.1 Sectors . . . 39
3.2 401-1a: Left: <30; Center:30-50; Right:>50 . . . 43
3.3 401-1a. Sectors & MIB ESG boxplots . . . 43
3.4 401-1b: Left: D; Right:U . . . 44
3.5 401-1b. Sectors & MIB ESG boxplots . . . 44
3.6 401-1c. . . . 45
3.7 401-1c. Sectors & MIB ESG boxplots . . . 45
3.8 401-1d. . . . 46
3.9 401-1d. Sectors & MIB ESG boxplots . . . 46
3.10 403-9a. . . . 47
3.11 403-9b. . . . 48
3.12 404-1a. . . . 49
3.13 404-1a Sector & MIB ESG. . . . 49
3.14 404-1b Dirigenti. . . . 50
3.15 404-1b Quadri. . . . 51
3.16 404-1b Impiegati. . . . 51

3.17 404-1b Operai. 52
3.18 405-1a. 53
3.19 405-1b. 54
3.20 405-1b. Sector & MIB ESG 54
3.21 405-1c. % F Tot 55
3.22 405-1c. % F Dir 55
3.23 405-1c. % F Qu 56
3.24 405-1c. % F Im 56
3.25 405-1c. % F Op 57
3.26 405-1d. Dir 57
3.27 405-1d. Dir Sector & MIB ESG 58
3.28 405-1d. Qu 58
3.29 405-1d. Qu Sector & MIB ESG 59
3.30 405-1d. Im 59
3.31 405-1d. Im Sector & MIB ESG 60
3.32 405-1d. Op 60
3.33 405-1d. Op Sector & MIB ESG 61
3.34 405-2. Dir 62
3.35 405-2. Qu 63
3.36 405-2. Im 63
3.37 405-2. Op 64

4.1 Euclidean dendograms 68
4.2 Manhattan dendograms 68
4.3 Corrplot: correlation among the 91 companies 69
4.4 Pearson correlation dendograms 70
4.5 Network 71
4.6 Dendogram cut at an height about 0.9 produce two different clusters . . . 72
4.7 $\mathcal{C}_1$ is the red partition; $\mathcal{C}_2$ is the golden one. 73
4.8 $\mathcal{C}_1$ is the green partition; $\mathcal{C}_2$ is the pink one. 74

List of Tables

3.1 401-1a 42
3.2 401-1b 43
3.3 401-1c 44
3.4 401-1d 45
3.5 403-9a 46
3.6 403-9b 47
3.7 404-1a 48
3.8 404-1b 50
3.9 405-1a 52
3.10 405-1b 53
3.11 405-1c 54
3.12 405-1d 57
3.13 405-1d 58
3.14 405-1d 59
3.15 405-1d 60
3.16 405-2 61

4.1 Mean of the obtained communities. Superscript 1: Hierarchical clustering; Superscript 2: Modularity optimization. 76