U.PORTO

FEP FACULDADE DE ECONOMIA UNIVERSIDADE DO PORTO

SUWAN - A SUPERVISED CLUSTERING ALGORITHM WITH ATTRIBUTED NETWORKS

Bárbara Monteiro Santos

Dissertation Plan

Master in Modeling, Data Analysis and Decision Support Systems

Supervised by

Pedro José Ramos Moreira de Campos

2021

Abstract

A new method of supervised clustering with attributed networks is proposed, SUWAN. The goal is to obtain class-uniform clusters, while minimizing the number of clusters. This method deals with representative-based supervised clustering, where a set of initial representatives is randomly chosen. By assigning each observation to the closest representative, clusters are obtained. With the new methodology, the way nodes are associated to clusters does not only depend on their network distance, but also on the distances between their attributes. This can be accomplished through a combination of weights between the matrix of distances between nodes and their attributes, when defining the clusters. Hence, the method considers both structural and compositional characteristics of the network. As a benchmark, we use the subgroup discovery on attributed network data. Subgroup discovery focuses on detecting subgroups described by specific patterns that are interesting with respect to some target concept and a set of explaining features. Therefore, interesting patterns among subgroups can be revealed, for example, by inductive and exploratory data analysis tasks that find relations between a dependent and independent variable, considering the compositional aspect of the networks. For this work, SD-Map, a fast algorithm for exhaustive subgroup discover, will be used to perform subgroup discovery on attributed networks. The proposed methodologies are applied to an inter-organizational network, denominated by EuroGroups Register, a central register that contains statistical information on companies from European countries, provided by Statistics Portugal.

Keywords: Supervised Clustering; Attributed Networks; Subgroup Discovery.

Resumo

Um novo método de *clustering* supervisionado com redes atribuídas é proposto, denominado por SUWAN. O objetivo é obter *clusters* homogéneos tendo em conta uma classe, minimizando ainda o número de *clusters* obtidos. Este método lida com *clustering* supervisionado baseado em representantes, onde um conjunto de representantes iniciais é escolhido aleatoriamente. Ao atribuir cada observação ao representante mais próximo, os *clusters* são obtidos. Com a nova metodologia, a forma como as observações são associados aos *clusters* não depende apenas da distância de rede, mas também das distâncias entre atributos. Esta técnica é realizada através de uma combinação de pesos entre a matriz de distâncias entre os nós e os atributos, no processo de formação dos *clusters*. Portanto, o método considera as características estruturais e composicionais da rede. Paralelamente, é realizada uma análise de outro método, denominado *subgroup discovery*. O método de *subgroup discovery* foca-se na deteção de subgrupos, descritos por padrões específicos que são interessantes com relação a um determinado *target* e um conjunto de medidas explicativas. Deste modo, padrões interessantes entre subgrupos podem ser revelados, por exemplo, por tarefas de análise de dados indutiva e exploratória, que encontram relações entre uma variável dependente e independente, considerando o aspeto composicional das redes. Para este trabalho, *SD-Map*, um algoritmo rápido para descoberta exaustiva de subgrupos, é usado para realizar *subgroup discovery* em redes atribuídas. As metodologias propostas são aplicadas a uma rede inter-organizacional, denominada por *EuroGroups Register*, um registo central que contém informação estatística sobre empresas de países europeus, disponibilizado pelo Instituto Nacional de Estatística.

Palavras-chave: *Clustering* Supervisionado; Redes Atribuídas; *Subgroup Discovery*.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Many real-world interactions now generate data in a network structure, with connections among elements. These networks can assume different types, from social networks, such as friendship ties, to networks formed between objects, that are not social (Tabassum et al. 2018). One of the distinguishing characteristics of social networks is their propensity for displaying community structure, which means that, groups of densely linked vertices that are poorly connected to other groups of vertices may be revealed (Oliveira and Gama, 2012). According to Harenberg et al. (2014), community detection consists of detecting groups of densely connected nodes that typically have fewer connections to nodes outside that group.

Therefore, the aim of community detection algorithms is to find cohesive subgraphs of nodes that can be representative of a community, focusing on the structural aspects of the network. A remarkable contribution to the field was made by Newman, that developed an algorithm that measures the quality over the possible divisions of a network, also known as Modularity (Newman, 2006). On the other hand, stands a clustering approach, that allows studying a set of elements by splitting it into smaller groups with similar characteristics, with focus on the compositional characteristics of the network. For this approach, two main methodologies are highlighted, hierarchical and partitional clustering algorithms. Partitional clustering methods discover all clusters concurrently and do not enforce a hierarchical structure, whereas hierarchical clustering algorithms find nested partitions iteratively (Jain, 2010).

In this work, a new methodology of supervised clustering, that considers both structural and compositional characteristics of the network, is developed. The new method SUWAN, Supervised clustering With Attributed Networks, is proposed, based on the Single Representative Insertion/Deletion Hill Climbing with Restart (SRIDHCR) algorithm from Eick and Zeidat (2004). As a benchmark, subgroup discovery is used to detect and identify relevant network patterns (Helal 2016). The goal is to discover interesting associations among different variables with respect to a property of interest. In contrast with standard community detection methods, SUWAN groups elements of a graph, based on their structural and compositional characteristics, while it provides class-uniform clusters, based on a predefined target variable. Subgroup discovery on the other hand, can

provide a description and identification of communities based on the combination of their features.

On the field of supervised clustering, some authors argue that classical techniques of clustering do not guarantee that objects of the same class are grouped together (Al-Harbi and Rayward-Smith, 2006). The proposed method can deal with this limitation and improve clusters purity. Furthermore, it tackles the unexplored field of supervised clustering on attributed networks. Alternatively, other authors claim that classical community detection techniques focus only on finding subgroup of nodes with a dense structure, lacking an interpretable description. Therefore, subgroup discovery can deal with description-orientated community detection. Moreover, this approach can also provide insights beyond connectivity within communities, and the relationships between subgroups of nodes as well.

An application of the new method is made on the dynamic of networks in business. The inter-organizational network in study, denominated by EuroGroups Register (EGR), holds relevant statistical information about organizations and the aim is to find patterns on this business register. The dataset contains over 6870 networks for the most recent year of 2018. For this purpose, social network analysis, graph theory, supervised clustering, and subgroup discovery will be applied to extract useful information, in order to discover communities among networks, based on a target variable, that measures the business performance.

Business performance from a network perspective is an increasing area of study for economists and social scientists. Organizations are in a constant adapting process to the changes of the environment (Lechener & Dowling 2003). Every day new firms are brought to the surface, and some others just disappear into another company. From a social network point of view, the relations between organizations can be perceived as a graph, where nodes are representative of companies and their relationship is represented by links. The merging and acquisition of firms are an ongoing process for larger companies that aim to develop their business. For some cases, a simple join of forces can also be accomplished through collaborative strategies that can assume the form of joint ventures, strategic alliances, holdings, among others. Hence, organizational networks aim to improve performance (Hoberecht et al., 2011). This way, organizations tend to be more connected among them, establishing their networking with other organizations.

Software R, version 4.0.2, was used for the experimental component of this study.

1.1 Motivation

The main focus of this dissertation is to propose a new supervised clustering algorithm for attributed networks. The need for such an approach started when we realized that it is important to cluster inter-organizational networks according to the structural position and also their own characteristics on the network. The employment of supervised clustering in attributed networks has not yet been applied. Therefore, the aim is to contemplate the structural and compositional characteristics of the network in the clustering process. Moreover, subgroup discovery on networks is an underdeveloped topic, and it has been studied by few authors like, Lucas et al. (2019), Deng et al. (2020) and Atzmueller & Mitzlaff (2011). Furthermore, the application of subgroup discovery on networks has some limitations, since it only focuses on finding the most interesting groups of nodes, leaving observations with no defined subgroup. This way, a contribution of this work is the exploration of supervised clustering using a benchmark with subgroup discovery on inter-organizational networks.

Additionally, this work is being developed with a direct cooperation of Statistics Portugal (INE), that provides the database in analysis. In fact, there is a clear loophole related with data of organizations networks since this information is not easily accessed. Given that, the opportunity to explore an inter-organizational network from real-world data can provide some useful insights on how organizations are structured and connected between themselves and what impacts their performance. Furthermore, the database provided by INE has never been explored in this perspective and, therefore a brand-new input on a network analysis for organizations can enable new discoveries and set paths for future analysis.

1.2 Organization of dissertation

This dissertation is organized in five chapters. The first chapter introduces the topics studied in this work and the motivation for this work.

In Chapter 2, is presented the literature review on the main topic. This way, this chapter starts by introducing some basic concepts on the topics of inter-organizational networks, social network analysis, attributed networks, and community detection. Then, on the last section of this section, a more detailed literature review is made on subgroup discovery and supervised clustering.

In Chapter 3, the focus is on methodology and data. The first part introduces the research questions of the thesis, followed by a detailed description of the new methodology to be implemented. Next, a description on how the performance of both methods can be compared is presented, through a new evaluation measure, based on the clustering purity. At last, the EuroGroups Register data base is presented, describing all the important features for analysis.

In Chapter 4, the methodologies are applied to the EGR data set. The performance of both methods is compared, while an interpretation of the output is made separately for the two methodologies.

In Chapter 5, an analysis on the variables that have more impact on the organization's performance is presented. Moreover, it is analysed the impact of the network topology on the business performance. For that, a correlation analysis between network topology measures and the organizational turnover is presented, followed by a multiple linear regression model.

Finally, Chapter 6 presents the conclusions and challenges of the work developed, presenting some final remarks and limitations of the implemented methods.

Chapter 2

Literature Review

2.1 Inter-Organizational Networks

Nowadays, the organizational structure of firms is in constant change. Organizations tend to adapt and change in order to gain a competitive advantage. Nevertheless, firms can also accomplish their goals through collaboration with other organizations. An increasing area of study for economists and sociologists is the varying organizational structures between business networks. Whether they are defined as strategic alliances, trade networks, joint ventures, or considered to be a result of the nature of the industry or local circumstances, they are seen as a mode of economic cooperation (Ebers, 1999).

According to Ricciardi and Rossignoli (2015), organizations are influenced by their inter-organizational relationships. With this, the existing relationships among organizations can be represented by a network. Thus, the network concept can also be applied to organizational structures. An inter-organizational network represents the relationships between different organizations, where organizations are represented as vertices, and their relationships by edges.

Hoberecht et al. (2011) state that organizational networks aim to improve performance. According to the authors, there are many reasons for establishing a network. For instance, when organizations want to achieve a specific goal that is shared by another organization. Moreover, to maximize supply chain efficiency and profitability, the corporate community invests in inter-organizational networks. Matous and Todo (2017), studied the impact of the network topology and diffusion on Japanese automobile production networks, that reveal the reorganization of inter-organizational networks and the organization performance coevolution. In a complementary line of thought, according to Popp, Milward et al. (2014), there are three types and functions of networks between organization – information diffusion and knowledge exchange, network learning and innovation.

The evolution of inter-organizational networks may be perceived as a cyclical cooperation system (Greiner & Levati, 2005). However, networks can still develop over time, going through four stages of formation, development and growth, maturity, sustainability, and resilience and, finally death and transformation (Popp et al., 2014).

As a method of evaluating networks, social network analysis remains highly useful, particularly as a way to understand the nature and content of relationships of different types (Popp et al., 2014).

2.2 Social Network Analysis and Graph Theory

Social network analysis (SNA) emerged in the field of social science, and it was used to manage theoretical questions and problems, trying to explain the social behavior by means of the net-work structure of societies. Indeed, SNA seeks to uncover patterns withing the relationships of certain groups, usually, links between human beings, but not necessarily.

In fact, L. Freeman (2004) states that the social relationships may concern other types of connections besides humans, such as animals or even organizations. Social network analysis can be perceived as an interdisciplinary field, since that, apart from social science, it also established an important role for the fields of biology, business, computer science, among others.

Wasserman and Faust (1994) define social network as a finite set or sets of actors and their specified relationship or relationships. Hence, for a social network, there must be social entities, also referred as actors, and a relational tie, that links two or more actors. Typically, this network is represented by a graph, with vertices and edges as the two fundamental elements.

Different real-world connections can be graphically represented by a set of points and their connections. Graphs are therefore a way to map social structures, that can assume different forms. In the field of social relations, graphs can show friendship ties between actors (Ball & Newman, 2013). On the other hand, it can be representative of an information networks (Harvey, Kleinberg, & Lehman, 2006) or even biological networks (Alon, 2003).

With this, a graph G can be defined as a nonempty set of vertices, V, and a set of edges, E, denotated by $G(V, E)$. Besides the graphical representation, there are two traditional ways of representing a graph G, using an adjacency matrix or list. An adjacency list consists in the representation of all edges in graph as a list, while an adjacency matrix is a representation of which vertices are adjacent to which other vertices (Singh & Sharma, 2012). In the matrix structure, all vertices of a graph are displayed, and, when dealing with unweighted networks, the matrix can be filled in a binary system, with zeros and ones, indicating the absence or presence of a connection between two vertices, respectively. In

the case of weighted networks, the matrix is field with the correspondent weights, where zero represents a non-existent connection.

Another distinction can be made between direct and undirect graphs. For undirected graphs, a pair of vertices is either connected or not, while in directed graphs, the connection can assume a direction, either by a single link or a double link (Palla, Farkas, Pollner, Derényi, & Vicsek, 2007).

2.3 Attributed Networks and Community Detection

Beyond the structural form, where nodes engage on relationships, defined by links, networks may contain additional information, which can be related to the entities and their relationships. Therefore, adding these additional feature data to the corresponding nodes and/or edges generates an attributed graph (Hewapathirana, 2019). On attributed networks, nodes and/or edges are labeled with additional information, allowing further dimensions for detecting patterns that describe a specific subset of nodes of the network.

In the context of attributed networks, some notions can be formalized as the following. Let $G = (V, E, A)$ be the denotation for an attributed graph, with n set number of vertices (V), m number of edges (E) and A set of attributes. For each attribute a_k, a range $dom(a_k)$ of values is defined, such that, $a_k (v_i) \in A$, $v \in dom(a_k)$ and $v_i \in V$.

Community detection, as one social network analysis method, aims to detect subgroups of individuals that are densely, or cohesively, connected by a set of links. For detection of those cohesive subgroups and communities, some methods can be applied, such as, hierarchical clustering, with the single-link and complete-link algorithms, or clique-based methods that find fully connected subgroups, known as Cliques. These approaches are based on the structure of the network, namely how nodes are linked among them.

Fortunato and Hric (2016) argue that the identification of communities may provide insight into how the network is structured. Therefore, it can help to define vertices based on their position in relation to the communities to which they belong. Moreover, the authors refer that community detection in graphs can identify modules and, perhaps, an hierarchical organization, based on the graph topology. Nevertheless, the structure of a social network may not be enough to identify its communities (L. C. Freeman, 1996).

In the past years, numerous algorithms were proposed to solve the issue of community detection in attributed networks. Therefore, community detection is applied to find communities, such that vertices in the same community are densely connected in the graph

and have similar attribute values. Among others, algorithms such as I-Louvain, ANCA and SAC2 can be highlighted for this task. The I-Louvain algorithm focus on the optimization of a global criterion in order to evaluate the similarity between the vertices' attribute values (Combe et al., 2015). On the other hand, ANCA – Attributed Network Clustering Algorithm (Falih et al., 2018), emphasizes the topological information of the network over a set of new features. Lastly, the SAC2 algorithm (Dang & Viennet 2012) relies on the structure of the graph, as well as the similarity between nodes' attributes.

2.4 Subgroup Discovery

Subgroup discovery (SD) is a data mining technique that focus on discovering interesting relationships between different objects (Herrera, Carmona, González, & Del Jesus, 2011). In fact, SD is not applied to find all the possible subgroups, but rather to find the best one, thus, most interesting, or unusual subgroups (Wrobel, 1997).

One main advantage of SD is the ability to deal with real-world data, i.e., characterized by its large size and complexity, involving many attributes and different data types (Meeng & Knobbe, 2020). Therefore, SD is broadly applied to real world problems in the areas of Health, with detection of risk groups diseases diagnosis, such as cancer (Mueller et al., 2009), Marketing (Gamberger & Lavrac, 2002), E-learning (Carmona, González, Del Jesus, Romero, & Ventura, 2010) and Spatial subgroup mining, applied for example to demographic area (Andrienko, Andrienko, Savinov, Voss, & Wettschereck, 2001).

According to Atzmueller and Puppe (2006), a subgroup discovery setting depends on four main properties: (*i*) the target, (*ii*) the subgroup description language, (*iii*) the quality function and (*iv*) the search strategy. The description language specifies the individuals that belong to the subgroup. Therefore, a subgroup description can be defined by the conjunction of a set of selection expressions that are part of the attribute's domain. To measure the interestingness of a subgroup, a quality function can be set based on a statistical evaluation function. The search method is used to rank the subgroups discovered when searching. The disparity in the distribution of the target variable concerning the subgroup, the general population, and the size of the subgroup are common quality parameters. With this, for a particular target variable, a quality function is used to evaluate a subgroup description and to rank the discovered subgroups during the search.

Different algorithms have been developed and applied to subgroup discovery task (Table 2.1). The pioneers in this field were EXPLORA (Klösgen, 1996) and MIDOS

(Wrobel, 1997). Both algorithms are extensions of classification algorithms and use decision trees as the base method. An extension of these two previous algorithms is presented by Klösgen and May (2002), with the SubgroupMiner algorithm, that combines decision rules with an interactive search in the space of solutions. Another method developed in the extension of classification algorithms is CN2-SD (Lavrač, Kavšek, Flach, & Todorovski, 2004), adapted from standard classification rule learning approach CN2 to subgroup discovery.

In association rule algorithms, the aim is to obtain relations between the variables, generating rules that can have variables both in the antecedent and consequent form. In subgroup discovery, the consequent of the rule consists of the property of interest, that is prefixed (Herrera et al., 2011). Thus, this characteristic makes it feasible to adapt association rule algorithms to subgroup discovery.

Atzmueller and Puppe (2006) propose a fast algorithm for exhaustive subgroup discovery, SD-MAP, based on FP-growth algorithm for mining association rules adaptation for SD. In the same way, APRIORY-SD algorithm was developed by adapting association rule learning to SD task. This method was built under the classification rule learner APRIORY-C, using a weighted scheme in rule post-processing, weighted relative accuracy, as a quality measure, and a probabilistic classification of instances (Kavšek & Lavrac, 2006).

Table 2.1 - Examples of subgroup discovery algorithms according to theirs field of extension.

Field of Extension	Subgroup Discovery Algorithms
Classification	EXPLORA
	MIDOS
	SubgroupMiner
	CN2-SD
Association	SD-MAP
	APRIORY-SD

2.4.1 Subgroup Discovery in Networks

The need to discover and analyze interesting patterns in data also emerged around networks, being social networks the most commonly studied.

Lucas, Gomes, Vimieiro, Prudêncio, and Ludermir (2019) investigated how the problem of group profiling can be modeled as a subgroup discovery task. For group profiling, on traditional univariate methods, a relevance function is needed, that measures the importance of each feature to distinguish the members of a community. This method

evaluates features as independent, neglecting possible interesting interactions that could enhance the overall description of a community. With this in mind, the authors propose that group profiling would benefit on a multivariate approach, which accounts interactions between features and could return the best subsets of features describing a community. Another characteristic of this method is the incorporation of coverage of a description. Descriptions with low coverage only represent a small part of the members of a community, therefore, communities may be described by an incomplete or by a pattern that occurred by change, that would not describe the entire community. Subgroup discovery is used to get a more expressive and comprehensive description that allow for a better understanding of groups as a whole.

Deng, Kang, Lijffijt, and Bie (2020) developed a work on graph mining with a subgroup discovery approach. According to the authors, the connectivity structure of a network is related with the attributes of the nodes, therefore, it can be understood in terms of patterns of subgroups of individuals with certain properties that are differentiated from other subgroups of individuals. Therefore, a method that incorporates an interestingness measure was proposed, in order to find pairs of node subgroups which the edge density is considered to be interestingly high (or low). This method provides an improvement over interestingness measures used on subgroup discovery for dense subgraph mining in attributed graphs. The main idea of this method was to overcome the classical community detection that typically assumes that links only exist because nodes share similar attributes, which limits sparce subgraph. Moreover, it extends the research beyond connectivity within communities, analyzing the relations between subgroup of nodes.

Atzmueller and Mitzlaff (2011) proposed an efficient descriptive community mining using subgroup discovery as a pattern mining technique. The authors propose a method that collects patterns that describe communities by combination of features. This way, they are able to identify and describe interesting communities, in contrast with standard community mining approaches that only focus on identifying communities as subsets of users. The graphs presented by the BibSonomy system show explicit friendship relations between users, therefore they directly indicate communities according to the link structure of the network. Additionally, communities of users can be characterized in terms of their descriptive features, such as, bookmarks or publications, that can be used as tags. For the subgroup discovery, a quality function is defined as an optimistic estimation of a subgroup. The proposed method extends the use of optimistic estimates for efficiently searching the

description space while optimizing the community measures on the network structure at the same time. For this purpose, local modularity and inverse conductance were used as optimistic estimates for community mining. For the subgroup discovery task, COMODO algorithm was applied for mining community patterns.

Atzmueller, Doerfel, and Mitzlaff (2016) make a comparison between classical community mining methods and subgroup discovery, arguing that community detection only identifies subgroup of nodes with a dense structure, lacking an interpretable description. Usually, community mining methods only consider the nodes of a network as mere strings or ids, and do not provide an intuitive description of the community, for example, an easily interpretable conjunction of attribute-value pairs. The aim is to identify communities as sets of nodes together with a description, for example, with a logical formula on the values of the node's descriptive features. Therefore, the focus is to present description-oriented community detection. For this, an adapted subgroup discovery approach was used, with COMODO algorithm (Atzmueller et al. 2016). This method deals with, for example, small community sizes, that it is not addressed by standard approaches for community detection. Moreover, instead of finding a complete and global partitioning of the network, it considers a set of local, potentially overlapping communities. This method is based on the assumptions that social graphs are statical and not dynamic, overlapping communities are assumed possible, since entities in a network tend to belong to different communities, and the focus of the discovery is on local communities.

Atzmueller (2018) formalizes the problem of detecting compositional patterns in attributed networks, capturing dyadic subgroups that have a relevant behavior, based on the quality measure. According to the author, typical approaches for community detection and graph clustering only focus on the structural information of the network. This work allows a compositional analysis of attributed networks by exploiting the attribute information. A subgroup discovery and exceptional model mining techniques are put in use to detect patterns in subgroups of nodes that show an interesting and/or unusual behavior. Unlike classical community detection approaches, the focus is not on the structural aspect of the graph, but rather on the described features of the nodes contain in the network. The task is to identify communities as sets of densely connected nodes together with a description on the node's features. For this matter, an interestingness measure was defined based on two properties, duration and frequency of the interactions.

Table 2.2 - Summary of literature review of subgroup discovery on networks.

Authors	Objectives	Methodology	Main Conclusions
Lucas et al. (2019)	Develop a new method for Group profiling based on Subgroup Discovery	Perform a multivariate analysis that could capture the interactions among features. Incorporate a coverage of descriptions into the community detection.	Test performance on a real-world data set of scientific articles from Arxiv, to obtain descriptions for communities of authors in the co-authorship network of articles. Performance evaluation compared to the univariate strategy shows the compromise between quality and coverage of descriptions.
Deng et al. (2020)	Propose a method that finds pairs of node subgroups between which the edge density is interestingly high or low, using an information-theoretic definition of interestingness.	Formalize a subjective interestingness measure that allows to find patterns that describe the graph density between a pair of subgroups.	This method was able to identify patterns that provide genuine insight into the high-level network's structure based, identifying not only dense but also sparce subgraphs and describing the density between subgroups.
Atzmueller and Mitzlaff (2011)	Mining descriptive patterns in communities	Pattern mining using subgroup discovery. Maximizing local quality function for single communities. Optimistic estimates for efficient knowledge discovery in the context of subgroup discovery.	Application on data from the social bookmarking system BibSonomy. The results show a reduction on the search space based on optimistic functions used to perform the mining descriptive community patterns.
Atzmueller et al. (2016)	Develop a description-oriented community detection based on subgroup discovery	Provide structurally valid and interpretable communities using descriptive features of graph's nodes. COMODO algorithm for obtaining the k-best community patterns based on a community evaluation measure	Application on real-world data social systems. The algorithm was able to detect communities that are typically captured by shorted descriptions, leading to a lower description complexity. The results indicate statistically valid and significant outcomes that don't encounter typical problems such as small community sizes.
Atzmueller (2018)	Apply a compositional perspective for identifying compositional subgroups patterns on attributed social interaction networks	Adapt principles of subgroup discovery to the dyadic network setting. Detecting compositional patterns and capturing subgroups of nodes that show an interesting behavior according to their dyadic structure, estimated by a quality measure.	Application on a social interaction networks, captured at two conferences. The results indicate interesting findings according to common principles observed in social interaction networks, such as, the influence of homophilic features on the interaction. Moreover, the quality function allows to focus on specific properties of interest, in both simple attributed networks or multigraph representations.

2.4.2 Subgroup Discovery in Attributed Networks

In contrast with standard community detection methods, that only focus on the structure of subgroups and communities, subgroup discovery focuses on detecting subgroups described by specific patterns that are interesting with respect to some target concept and a set of explaining features. For this matter, subgroup discovery aim is to revel interesting patterns among subgroups, for example, by inductive and exploratory data analysis tasks that find relations between a dependent and (several) independent variables (Atzmueller, 2015), considering the compositional aspect of the networks.

In order to define subgroups, two criteria's can be taken into account, compositional and structural. Compositional criteria respects to the node attributes while structure criteria focus on tie structures. Cohesive subgroups on social networks may reveal the most involved participants of a community (Chin, Chignell, & Wang, 2010). This way, with the additional information supplied by attributed networks, subgroup discovery method can be applied in order to combine both structural and compositional characteristics of the network.

As previously referred, different types of algorithms can be used on a subgroup discovery task. However, some of them can be adapted to perform SD on networks. In fact, Atzmueller and Lemmerich (2009) first proposed a subgroup discovery algorithm, COMODO, that is an adaptation of SD-MAP algorithm. The developed work uses the SD-MAP algorithm with the preprocessing of the COMODO algorithm. A description on both methods follows.

2.4.3 SD-MAP Algorithm

SD-Map is an exhaustive subgroup discovery algorithm that consists in an adaptation of Frequent Pattern Growth method (FP-growth) for the subgroup discovery task. The FP-growth algorithm uses a compressed representation of the itemset database. The tree growths by tracking each itemset and mapping it to a path. This method has proven to be quite efficient, and it is based on a divide-and-conquer frequent pattern growth.

This way, SD-Map uses a modified FP-growth step that can explicitly compute the quality of the subgroup without referring to other intermediate outcomes (Herrera et al., 2011). Therefore, an extended prefix-tree-structure is used to store information for patter refinement and evaluation, hence, an extending FP-tree structure can be developed towards a single and multi-target concepts.

This method can deal with both binary and continuous targets. An FP-tree node stores the subgroup size and the true positive count of the respective subgroup definition for the binary case, and it considers the number of values of the target variable in the continuous case, allowing to determine the respective value of the quality functions accordingly.

As the associations measure the confidence and the support of rules, for subgroup discovery, a special quality function is used to measure the interestingness of a subgroup. Atzmueller and Puppe (2006) defines the subgroup quality computation based on the true positives *tp* (cases containing the target variable *t* in the given subgroup *s*), the false positives *fp* (cases not containing the target *t* in the subgroup *s*) and the positives and negatives regarding the target variable *t* in the population size *N*. A quality function is used to assess a subgroup description and rate the identified subgroups during search (Atzmueller and Lemmerich, 2009). Examples for quality functions are given by Equations 2.1, 2.2 and 2.3, where *n* indicates the size of the subgroup, while *p* and *p0* are the relative frequencies in the target variable and total population, respectively.

$$q_{WRACC} = \frac{n}{N}(p - p_0) \quad \textbf{(2.1)}$$

$$q_{PS} = n(p - p_0) \quad \textbf{(2.2)}$$

$$q_{LIFT} = \frac{p}{p_0} \quad \textbf{(2.3)}$$

2.4.4 COMODO Algorithm

The SD-MAP algorithm for fast exhaustive subgroup discovery, and the FP-growth algorithm for mining association rules form the basis of COMODO algorithm. This way, COMODO is an adaptation of the SD-MAP algorithm. This method focuses on description-oriented community detection, using subgroup discovery (Atzmueller, 2018). Using both structural and compositional characteristics of a graph, it is possible to identify communities as set of nodes together with a description.

COMODO algorithm is a fast branch-and-bound algorithm that relies on optimistic estimates. As referred, COMODO uses an extended FP-tree structure to efficiently navigate the solution space.

The inputs for this algorithm are the graph structure and the descriptive information of the attributed graph. Therefore, this information can be contained in two different data

structures, a graph structure encoded in a graph *G* and the attributed information contained in a database *D*. In a first step of preprocessing data, these two data sources are merged, creating a new transformed dataset where each data record represents an edge between two nodes. This method can be applied if there are no isolated nodes in the network, therefore nodes can be described as sets of edges. Each data record's attribute values are the common attributes of the edge's two nodes.

Afterwards, it is generated an initial community pattern tree (CP-tree), where each CP-node of the CP-tree captures information about the aggregated edge information concerning the dataset and the respective graph.

The result of the COMODO algorithm is then the set of the top-k community patterns.

2.5 Supervised Clustering

Clustering is a methodology consists of grouping data according to a desired criterion, allowing finding a structure in a dataset (Sinaga & Yang, 2020). This way, the main goal of clustering analysis is to classify a set of items into homogenous groups, also referred as clusters, with a pre-determined measure of similarity. When clustering is performed, the similarity measure should be higher within groups, when comparing to the similarity between different groups (Jain, Murty, & Flynn, 1999).

The problem of supervised clustering may be presented as a pair (X, C), where X denotates a limited set of items $X=\{x_1,\ldots,x_n\}$ and C denotates a group of distinct and nonempty subsets $C_1,\ldots,C_k$ of X. Similarly to traditional clustering, a distance function can be formalized as $dist(x,y)$, where x and y represent two distinct data points. Depending on the type of data, different functions can be used to measure the distance between data points. For numerical data, the most commonly used are Euclidean distance (Equation 2.4) and Manhattan distance (Equation 2.5). On the other hand, for nominal attributes, the most used measure is based on a simple matching method, denominated by Hamming distance (Equation 2.6), where $dH(x,y)$ corresponds to the number of places where x and y are different (Pandit & Gupta, 2011).

$$dE(x,y)=\sqrt{(x-y)^2} \tag{2.4}$$

$$dM(x,y)=|(x-y)^2| \tag{2.5}$$

$$dH(x,y)=\left|\{i:x_i\neq y_i\}\right| \quad \textbf{(2.6)}$$

Unlike traditional clustering methodology that works around non-labeled data, the assumption in supervised clustering is that the items are classified. The objective of supervised clustering is to find class-uniform groups of items,that have a high probability density with respect to a single class (Eick & Zeidat, 2004). Next, a summary of the work developed on supervised clustering is presented.

Eick and Zeidat (2004) present a k-medoid-style clustering algorithms for supervised clustering. The k-medoid model aims to search for *k* representative objects, known as medoids, that reduce the average dissimilarity of all the data set's objects to the closest medoid (Kaufman & Rousseeuw, 1987). This way, the clusters are obtained based on the group of objects that have been assigned to the same medoid. One of the proposed algorithms, based on k-medoid model, is Supervised Partitioning Around Medoids (SPAM). This technique uses a fitness function, instead of the dissimilarity measure, and the number of clusters *k*, as an input parameter. The fitness function is presented in Equation 2.7, where *n* represents the total number of examples, *c* the number of classes and β defines the penalty associated with the number of clusters *k*. The class impurity measures the percentage of minority examples in the different clusters of a certain clustering. The goal is to minimize *q(X)*, to obtain class-uniform clusters, while minimizing the number of associated clusters.

$$q(X)=\text{Impurity}(X)+\beta\text{Penalty}(k) \quad \textbf{(2.7)}$$

$$\text{Where Impurity}(X)=\frac{\#\text{ of Minority Examples}}{n}\text{ and Penalty}(k)=\begin{cases}\sqrt{\frac{k-c}{n}}, & k\geq c\\ 0, & k<c\end{cases}$$

This way, the algorithm builds an initial solution, using as representative objects, the members of the most frequent class in the data set. Then, the algorithm repeatedly and greedily inserts non-representative objects to the current set of representatives that yields the lowest value for the fitness function *q(X)*. Another algorithm proposed by the authors is the Single Representative Insertion/Deletion Steepest Decent Hill Climbing with Randomized Starting (SRIDHCR). In the same way as SPAM, this method tries to minimize the quality function *q(X)*. It starts by randomly selecting a number of objects as an initial set of representatives. Then, by inserting and deleting single items from the

existing collection of cluster representatives, it greedily seeks for solutions. The major difference in this algorithm relies on the fact that the *k* number of clusters is not fixed, as the algorithm searches for an optimized value of clusters.

Similar work was developed by Gan et al. (2018), that employed a novel graph-based classification method, called Supervised clustering-based Regularized Least Squares Classification (SuperRLSC). The proposed methodology is based on the idea that supervised clustering may uncover more genuine data structures, when compared with the traditional clustering. This way, a supervised k-means algorithm is used, in order to partition the dataset into different meaningful clusters. Similar to Eick and Zeidat (2004), a fitness function is designed, as described in Equation 2.8, that allows to find as many homogeneous clusters as possible, while minimizing the number of clusters. The optimal solution is then obtained by computing the values of *J(X,M,β)*, with different number of cluster, *M*, and picking the ideal number of clusters that minimizes that fitness function.

$$J(X,M,\beta) = -\text{Purity}(X) + \beta\text{Penalty}(M) \quad \textbf{(2.8)}$$

Finley and Joachims (2008), present a method of supervised clustering with k-means algorithm. This work is based on the idea that, in order to successfully implement k-means, a similarity measure that reflects the properties of the cluster must be chosen prudently. Hence, a structural Support Vector Machines (SSVM) method is implemented to perform the k-means algorithm as a supervised task. This methodology uses a SSVM approach to learn a parameterized distance measure such that k-means may provide the preferred clusters and maximize the cluster accuracy. This way, the similarity measure is learned through given training examples of item sets with proper clustering so that future sets of items are grouped similarly.

In the same line of thought, Al-Harbi and Rayward-Smith (2006), propose a new method of adapting k-means for supervised clustering. The authors argue that the traditional k-means algorithm for unsupervised clustering does not guarantee to group the same classes of objects together. Therefore, it proposes an adaptation of k-means, as a classifier clustering algorithm. The proposed method attempts to partition the objects that have the same label into the same cluster, by modifying strategic steps of the traditional k-means algorithm, namely, the Euclidean metric and the objective function. The Euclidean metric used on k-means (Equation 2.9) is transformed into a weighted Euclidean metric (Equation 2.10), partition the data according to the different labels, and assigning a greater

weight to a chosen field, which has a more significant relationship whit those class labels. The process of choosing the appropriate set of weights can be seen as an optimization problem, addressed by any metaheuristic technique. In this case, Simulated Annealing is used to find the best set of weights for the clustering problem. The goal is to make the k-means algorithm's divisions as confident as possible.

$$\delta(x,y)=\sqrt{\sum_{i=1}^{n}\left(x_i-y_i\right)^2} \quad \textbf{(2.9)}$$

$$\delta_w(x,y)=\sqrt{\sum_{i=1}^{n} w_i\left(x_i-y_i\right)^2} \quad \textbf{(2.10)}$$

Table 2.3 - Summary of literature review of supervised clustering.

Authors	Objectives	Methodology	Main Conclusions
Eick and Zeidat (2004)	Find the best k class-uniform clusters, while minimizing the number of associated clusters.	Representative-based supervised clustering with implementation of a quality function that measures the purity and number of clusters.	Supervised clustering allows background knowledge of data, that can be used to retrieve subclasses and enhance classification algorithms.
Gan et al. (2018)	Build graphs of data through supervised clustering to uncover the intrinsic patterns and discriminative information hidden in the data.	Supervised clustering-based Regularized Least Squares Classification with implementation of a quality function.	Application on several UCI datasets that demonstrated that the data structure revealed by supervised clustering aids in the construction of more effective graphs, and the technique employed outperforms both traditional graph-based and state-of-the-art supervised classification methods.
Finley and Joachims (2008)	Train k-means unsupervised clustering by enhancing the partition of data and the similarity measure to obtain the desired clusterings.	Structural Support Vector Machines method implementation on k-means algorithm.	In comparison to naïve pairwise learning or unsupervised k-means, the proposed methodology performed better in experiments.
Al-Harbi and Rayward-Smith (2006)	Adapting k-means algorithm for use as a classifier.	Supervised k-means employed with a combination of Simulated Annealing and weighted k-means algorithm.	Simulated Annealing is used to discover the best weights for fields, and the k-means algorithm is used to construct clusters using the appropriate weighted Euclidean metric; The method proved to be efficient in mixed data, that contains both numerical and categorical variables.

Chapter 3

Methodology, the SWAN Algorithm and Data

3.1 Research Questions

This study follows a methodological approach to answer the following research questions:

- Can networks be clustered based on their performance?
- How can the structural and compositional characteristics of the networks be contemplated in the clustering process?
- Can the methodologies be evaluated compared based on the cluster's purity?
- Which method produces higher quality class-uniform clusters/subgroups?
- Are there crucial attributes that have an impact on the organizational performance?
- Does the topology of the network have an impact on the business performance?

3.2 SUWAN – Supervised clustering algorithm With Attributed Networks

SUWAN is a supervised algorithm for attributed networks. This new methodology employs SRIDHCR algorithm (Eick & Zeidat, 2004) that consists of representative-based supervised clustering with the addition of a quality function that assesses cluster purity and quantity. SUAWN allows clustering groups of nodes, considering the structural and compositional characteristics of the network. Additionally, it is suitable for either categorical or numerical attributes.

For this purpose, a fitness function is defined, as described in Equation 3.1. The input parameters concern the target variable, the penalty, β, associated with the number of clusters, and the weight of the network distances, α. The number of classes, *t*, is established by the number of unique values that the target variable can assume. The target variable should be selected to reflect a characteristic of interest when forming clusters.

$$Q(x)=\text{Impurity}(x)+\beta\times\text{Penalty}(k) \quad \textbf{(3.1)}$$

The pseudocode for SUWAN algorithm is presented below. The first step of the algorithm refers to the application on attributed networks that considers both structural and compositional characteristics of the network. Thereby, two matrixes are computed, one

that measures the distances between nodes attributes, D_1, and other that measures the distance between all nodes in the network, D_2. With this, a weighted matrix is obtained, with a ponderation defined by α, that determines the weight given to the distances between the nodes on the network.

The SRIDHCR algorithm (Eick & Zeidat, 2004) is employed based on k-means, which uses Euclidean metric, consequently, it is suitable only for numeric variables. To work around this limitation, a new dissimilarity measure for categorical variables, named Hamming distance (Equation 2.6), is incorporated in SUWAN algorithm. This measure is used to obtain the distances between nodes attributes (D_1). Hence, SUWAN employs either k-modes or k-means algorithms, according to the type of data.

SUWAN PSEUDOCODE

Input

$A=\{A_{n\times m}\}$ //Dataset, with *n* nodes and *m* attributes
$E=E\subseteq\{(x,y)|(x,y)\in V^2$ and $x\neq y\}$ // Edge list, with V as the set of vertices
Target Variable
Beta $0<\beta\leq 2$ //Penalty
Alpha $0\leq\alpha\leq 1$ //Weight of network distances

Output

R= f: V → C, where V is the set of nodes and C the set of clusters
$C=\{C_1,..,C_K\}$ //Clusters

Algorithm

1. Calculate $D=\{D_{n\times n}\}$ //Weighted distance matrix
2. Repeat *r* times
   - *2.1. curr* = a randomly created set of representatives, $t+1\leq curr\leq 2t$
   - 2.2. WHILE NOT DONE DO
     - 2.2.1. Create new solutions *S* by adding a single non-representative to *curr* and by removing a single representative from *curr*
     - 2.2.2. Determine the element *s* in *S* for which *q(s)* is minimal; In case of tie chose first one
     - 2.2.3. IF *q(s)* < *q(curr)* THEN *curr* = *s*
     - 2.2.4. ELSE IF *q(s)* = *q(curr)* AND $|s|<|curr|$ THEN *curr* = *s*
     - 2.2.5. ELSE terminate and return *curr* as the solution for this run
   - 2.3. Assign remaining nodes to *curr* based on D.
3. Report the best out of the *r* solutions found.

As in the SRIDHCR algorithm, the proposed method starts by randomly selecting a set of initial representatives, denominated by *curr* (Step 4.1). The number of elements contain in this solution define the number of clusters *k*. As previously referred, the goal is to minimize Q*(x)*, to obtain class-uniform clusters, while minimizing the number of associated clusters. Therefore, *k* is not fixed, as the algorithm searches for an optimized number of clusters. By assigning each node to the closest representative with a weighted matrix,

clusters are obtained.

The algorithm then starts to generate new possible candidates, *s*, by adding and removing a single non-representative node from the current solution, keeping the solution that improves the quality function. The algorithm then terminates when Q*(x)* reaches an optimum. Nevertheless, the algorithm can still keep iterating while reducing the number of clusters and without improving Q*(x)* (Step 4.2.4).

Table 3.1 – Cluster representatives iteration process on SUWAN algorithm, with a toy example.

Iteration	Representatives	Q(x)
0	A, B, C, D, E	0.090
1	A, B, C, D, E, **F**	0.050
2	A, B, C, D, E, F, **G**	0.040
3	A, B, C, D, E, F, G, **H**	0.035
4	A, B, C, D, E, F, G, H, **I**	0.033
5	B, C, D, E, F, G, H, I	0.031
6	C, D, E, F, G, H, I	0.030
7	C, E, F, G, H, I	0.020

The process of adding and removing single non-representatives from the current set of cluster representatives can be observed in Table 3.1, where it demonstrates the optimization on the quality function. From the initial set of representatives randomly chosen, A, B, C, D and E (Iteration 0), the algorithm starts to add and/or remove representatives, saving the sets that produce a lower value of Q*(x)* for the first iteration, where it added element F. The algorithm then starts to iterate again from that current solution (Iteration 1), until it reaches the optimized solution at iteration 2. This process goes on to the point that, adding and/or deleting items from the current solution does not improve the value of Q*(x)*.

3.3 Evaluation Measures

In this section, we introduce some Evaluation Measures to measure the quality of the methodology implemented. Measuring the quality of the clustering it is an important step of the method's implementation since it enables the comparison with other procedures. However, evaluating the quality of a clustering is challenging, as the correct clusters are not known.

The implemented methodology works around labeled data. The cluster evaluation can be accomplished through the purity of the clusters, and therefore, a new measure of the overall quality based on the cluster's purity was computed to achieve the quality of the clustering.

Let the classes in the data set A be $T=(t_1,\ldots,t_i)$, and the number of clusters C be $C=\{C_1,..,C_k\}$. The clustering output is presented in a table format, with k lines and i columns, indicating the number of clusters and classes, respectively. For each cluster, the purity is determine as presented in Equation 3.1, where $PR_k(t_i)$ is the proportion of class t_i in cluster C_k.The overall quality measure is then computed by the total purity of the whole clustering, given by Equation 3.2, where $|C_k|$ is the total number of nodes in cluster k, and $|C|$ the total number of nodes of the network.

$$\text{Purity}(C_k) = \max_i (PR_k(t_i)) \tag{3.1}$$

$$\text{Purity}_{\text{total}}(C) = \sum_{k=1}^{i} \frac{|C_k|}{|C|} \times \text{Purity}(C_k) \tag{3.2}$$

3.4 Data: the EuroGroups Register

The European Union's Member States have embarked on a project to integrate and develop their national company registries for statistics reasons (Eurostat 2010). Eurostat coordinates this initiative, with goals set and success reported at the Business Registers — Statistical Units Working Group's yearly meetings. Other European nations, notably those in the European Free Trade Association (EFTA) and candidate countries, are welcome to participate in the initiative.

The EuroGroups Register (EGR) is a system of registers that includes a central register maintained by Eurostat as well as registers in each EU Member State and EFTA country. Information regarding international company groups is kept in the central registry. The central register keeps track of multinational corporations with statistically significant financial and non-financial transnational operations in at least one European country. The EGR database is composed by a several distinct groups of firms, that form multinational corporations. Those multinational groups correspond to networks, making the EGR database suitable for a network analysis.

The EGR network's goal is to maintain a comprehensive, accurate, consistent, and up-

to-date collection of connected and coordinated statistics registries, which provides compilers with a standard framework of multinational enterprise groups, both global and truncated national groupings, functioning in the EU and EFTA economies, as well as their constituent legal units and companies, and also ownership and control connections between legal units. Enterprise group structures are built by using control relationships (more than 50% of ownership) between two legal units. Enterprise groups are compiled by subsidiary-parent control relationships linked with direct or indirect control relationships.

In order to fully understand how the EGR network is formed, some basic concepts such as Multinational Enterprise Group, Legal Unit, Enterprise, Global Group Head and Global decision Center, must be disclosed. Therefore, a summary of concepts is presented below in Table 3.2, and a detailed description is followed.

Table 3.2 - Summary of basic concept of EGR database.

Concept	Abbreviation	Definition
Multinational Enterprise Group	MNE	Enterprise group that has at least two enterprises or legal units. Can be domestic or foreign controlled
Legal Unit	LEU	Individuals or institutions legally recognized by law or that are engaged in an economic activity
Enterprise	ENT	Legal Unit producing economic goods and services
Global Group Head	GGH	Parent legal unit of an enterprise that is not controlled by any other legal unit. Unit on top of the control chain of the group
Global Decision Center	GDC	Unit where strategic decisions are taken. The goal is to produce meaningful statistics
Ultimate Controlling Institutional Unit	UCI	

A multinational enterprise group is one that consists of at least two businesses or legal entities that are based in separate countries. These multinational groups can be differentiated as Domestic Multinational and Foreign Multinational, depending on if the country of the group is Portuguese or not.

In most cases, the Legal Unit (LEU) is documented in one or more administrative sources. Legal units are not always represented in the same way by the sources used for statistical business registrations. These units might differ across nations and between various sources within a country. As a result, the LEU is ineffective as a statistical unit, especially in international comparisons. The following characteristics define a LEU: (*i*) it owns things or assets, (*ii*) it has obligations, and (*iii*) it makes contracts. The legal unit is always the foundation for the statistical unit known as the "enterprise", either alone or in

conjunction with other legal units. LEUs are legal entities whose existence is recognized by law independent of the individuals or institutions that may own or be members of them, as well as natural persons who participate in economic activity on their own.

An Enterprise is a Legal Unit that produces economic products and services and has financial and investment decision-making autonomy, as well as power and responsibility for allocating resources for the creation of such goods and services. It may be engaged in more than one productive activity, and it can assume the form of a company, non-profit organization, or unincorporated business. The enterprise is the statistical unit at which information about its transactions is kept, including financial and balance-sheet accounts, and from which international transactions, an international investment position (if applicable), consolidated financial position, and net worth can be calculated.

The Global Group Head (GGH) is an enterprise group's parent legal entity that is not controlled by any other legal unit, either directly or indirectly. Global and domestic group heads can be recognized in multinational company groupings. The GGH is the multinational enterprise group's group head, whereas the domestic group head oversees the multinational enterprise group's abbreviated national portion. As a result, it symbolizes the unit at the top of the group's control chain.

The Global Decision Center (GDC) is the unit in charge of making strategic choices for a business group. This role is equivalent to UCI, which stands for Ultimate Controlling Institutional Unit. The UCI is an institutional unit that works its way up the chain of command of a foreign affiliate that is not controlled by another institutional unit. The goal of the definition of the UCI is to produce meaningful statistics. The *GDC=UCI* relation is the core requirement for EGR to serve the related statistical fields.

The EGR data base is accessed through relational tables, which hold information on the Multinational Enterprise Groups networks. As presented in Figure 3.1 below, each table contains a primary key that identifies the groups, legal units, and enterprises, with the attributes *GEG_EGR_ID*, *LEU_EGR_ID* and *ENT_EGR_ID*, accordingly. Also, these variables allow identifying the connections within tables, as secondary keys. This way, the network is established by affiliation relationships between organizations, that can be perceived as a "*parent*" and "*child*" relationship between organizations.

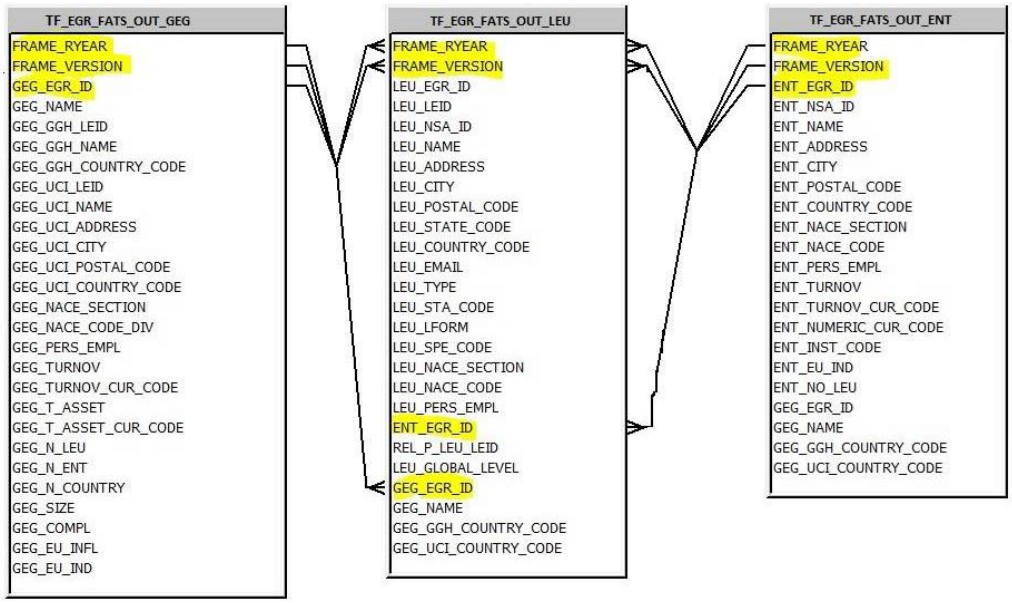


Figure 3.1 - EuroGroups Register database representation as relational tables between the groups, legal units, and enterprises.

For each main group presented in the first table of Figure 3.1, there is a parent legal unit that assumes the overall control, denominated as Global Group Head. This group detains a certain number of legal units and enterprises, identified by the attributes *GEG_N_LEU* and *GEG_N_ENT*, respectively. Figure 3.2 illustrates the graphical representation three different Global Group Head networks.

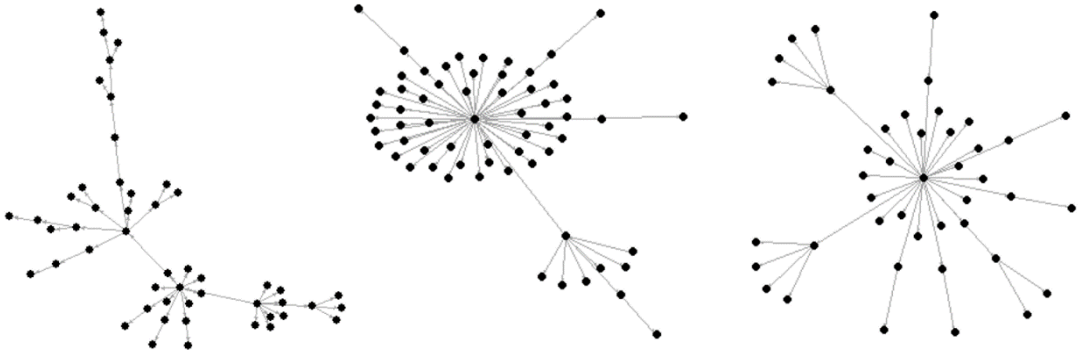


Figure 3.2 - Examples of networks topologies of different Group Heads and their Legal Units, with *igraph* package from R.

Among all variables available in the ERG database (Figure 3.1), only a few are relevant for the analysis performed. Therefore, in table 3.3 below, it is presented the list of attributes used to perform SUWAN and subgroup discovery on attributed networks, where some attributes are a result of a categorization.

Table 3.3 - Description of relevant attributes of EGR database.

Attribute	Description
LEU_LEID	ID of the Legal Unit
LEU_TYPE	List of type of Legal Unit (Brach or not)
LEU_LFORM	List of legal forms of Legal Units
LEU_COUNTRY_CODE	List of 2-digit ISO country codes
SIZE_CLASS	Size of the enterprise based on persons employed
TURNOVER_CLASS	Turnover class based on the enterprise turnover values
NACE_DIV	2-digit NACE Rev. 2 activity codes for the main activity of enterprises

A LEU can assume different forms, such as, limited liability company (LL), sole proprietor (SP), partnership (PA), government (GO), nonprofit body (NB), natural person (NP) or not defined (ND). The size of the Legal Unit is defined by the number of persons employed (*LEU_PERS_EMPL*), and it includes the total number of persons who work in the observation unit, as well as persons who work outside the unit but belong and are paid by it, such as sales representatives. Moreover, persons that are absent for a short period, on strike, part-time and seasonal workers, apprentices, and home workers on payroll are included in the counting.

The turnover class variable, groups into 6 classes the turnover values of each Enterprise. A Legal Unit may not have an associated turnover. On the other hand, an Enterprise has always an associated turnover, since that an Enterprise is a Legal Unit producing economic goods and services. These cases represent the Legal Units that are not Enterprises. From the network point of view, each LEU represents a node, so some nodes will not have a defined turnover class but still belong to the network structure. A more detailed description of the relevant attributes is presented in Annex A.

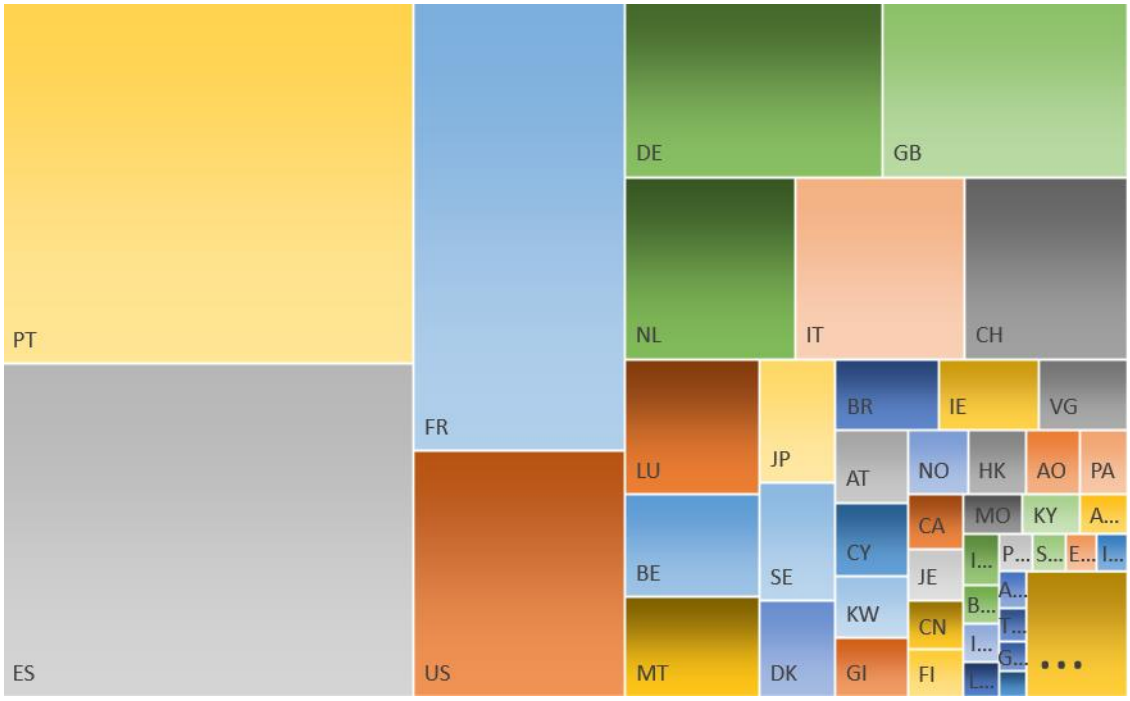


Figure 3.3 - TreeMap with hierarchical distribution of UCI countries codes for the year of 2018.

The groups covered by the ERG data base are based in different regions of the map, and the same group can hold companies from different countries. Despite that, the residence country of natural person that controls the group can be identified through the country code of the Ultimate Controlling Institutional Unit (UCI). For the year of 2018, there are over 90 countries covered by the EGR database. The distribution of the most frequent UCI countries is presented above, in Figure 3.3. It is noticeable that, more than half of the UCIs are settled in Portugal (PT), Spain (ES), France (FR) and United States (US), with number of UCIs per country of 1313, 1200, 836 and 458, accordingly. Other UCIs are spread amongst the remaining countries, with a smaller incision in countries like Andorra (AD), American Samoa (AS), Chile (CL), Romania (RO), among others.

Chapter 4

Results and Analysis

In this chapter, an application of the described methodology is performed on the EuroGroups Register. The available data for the EGR network, relates to the years 2016, 2017 and 2018. The analysis performed is made for the most recent year of 2018, that contains over 6870 groups of networks, even though not all of them are suitable for the analysis.

In order to present a detailed analysis of results, a subset of the total number of networks for the year of 2018 was created, based on the information retrieved from the TreeMap presented in Section 3.4. This way, the groups under analysis concern the networks which have an UCI based in Portugal. Moreover, the selected groups are also filtered by a minimum number of nodes of 20. In total, SUWAN and subgroup discovery were applied in 67 networks, that contain a total of 3848 LEUs. A list with the information about the networks under analysis is presented in Annex B, where the variable *GEG_N_LEU* indicated the number of nodes of each network. For confidentiality reasons, the networks are identified through numbering.

As previous described, clusters are obtained by associating nodes to the closest representative. Thus, the way clusters are formed depend on the weighted metric, that merges the weights of nodes attributes and network distances. This weight is defined by the parameter α, that sets the importance of the network distances, and its impact for clustering is illustrated in Table 4.1. It is noticeable that, when α is set to zero, clusters are formed based solely on the properties of the attributes, resulting in a significant dispersion between nodes from the same cluster on the network representation. On the opposite side, when only considering the networks distances, $\alpha=1$, nodes are grouped by closeness and clusters can be visually distinguished from each other.

To determine the optimized value for α, an analysis on the proportion of explained pseudo inertia was performed. For this, it was computed the partitions generated for a range of possible values of α, and a certain number of clusters k, of the distances between nodes attributes, D_1, and the distances between nodes D_2 (Chavent et al., 2018). To achieve this, a plot of the quality criterion Q_1 and Q_2 of the partitions P_k^α, obtained with varying values of α, was executed, as described in Equation 4.1, where $\beta=1$ represents the total proportion of the total pseudo inertia, based on the matrix of distances between the nodes

attributes (D_1), and $\beta=2$ represents the proportion of the total pseudo inertia, based on the matrix of distances between nodes (D_2). The higher the value of the criterion $Q_0(P_k^\alpha)$, the more homogenous the partition P_k^α is, from the node's attributes point of view. In the same way, the higher the value of the criterion $Q_1(P_k^\alpha)$, the more homogeneous the partition P_k^α from the node's distances point of view.

$$Q_\beta(P_k^\alpha)=1-\frac{W_\beta(P_k^\alpha)}{W_\beta(P_1)}\in[1,2] \qquad \textbf{(4.1)}$$

The optimized value of α is the trade-off point between the loss of nodes attributes distances and the gain of nodes distances. Figure 4.1 gives an example of a plot for the normalized proportion of explained pseudo-inertia calculated with the matrix of attributes distances (D_1) and the matrix of nodes distances (D_2).

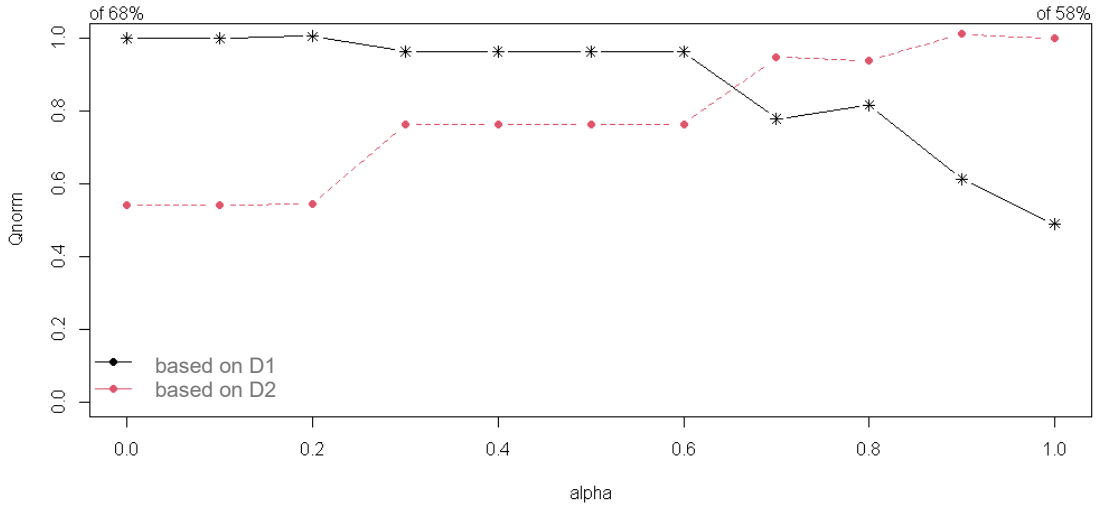
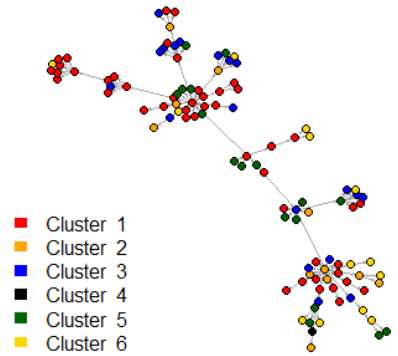
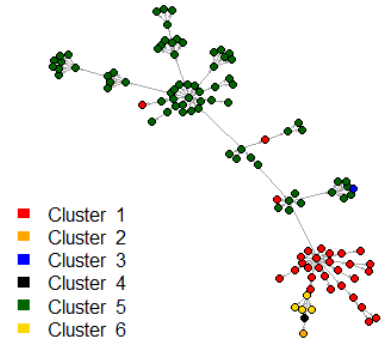
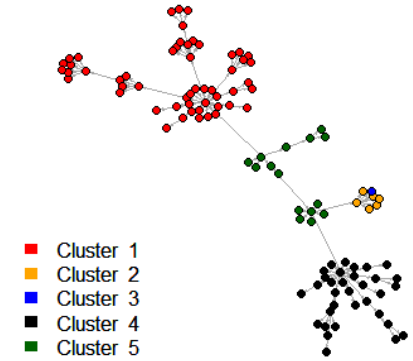


Figure 4.1 – Normalized proportion of explained pseudo-inertia for the distances between nodes attributes (D_1) and the distances between nodes (D_2).

For each one of the 67 networks under analysis, it was calculated the proportion of explained pseudo-inertia between D_1 and D_2, with α ranging between [0,1], and the number of clusters given by the number of class labels t, in the network, using *ClustGeo* package from R. The obtained average number of α on the 67 networks was 0.651. Therefore, the following analysis are performed with a weight of network distances of $\alpha=0.7$.

Table 4.1 - Impact of α variation on clustering analysis, with network representation.

Parameters	Network Representation
α = 0	
α = 0.7	
α = 1	

For both methods, SUWAN and subgroup discovery (SD), a target variable is defined to obtain the class labels. For the EGR network, the focus is to form cluster of enterprises that have the same turnover class.

Table 4.2 – Summary table of average performance of SUWAN and SD methods.

Algorithm	Average #Clusters/Subgroups	Average Overall Quality
SUWAN	3,299	0,532
SD	3,761	0,726

The results on the Portuguese UCI networks produced an average overall quality of 0.532 and 0.726 for the algorithm of SUWAN and SD, accordingly (Table 4.2). This means that,

on average, SD produced more purer clusters, when comparing with SUWAN. Concerning the number of clusters/subgroups produced, SD produces 3.8 subgroups, while SWUAN produced 3.3 clusters, on average. A table with the complete results per network on the overall quality of both methods is presented in Annex C.

4.1 Analysis of SUWAN results

Although SUWAN algorithm produced, on average, lower quality clusters, more that 11% of the networks achieved an overall quality higher than 70%. The results for these eight networks (38, 25, 48, 51, 32, 21, 41, and 50), are presented in Table 4.3 below.

Table 4.3 - SWUAN results on ERG network, with Portuguese UCI.

Network ID	#Clusters	#Nodes	Overall Quality
38	4	24	1
25	5	39	0,923
48	3	21	0,905
51	5	23	0,870
32	4	23	0,783
21	4	33	0,758
41	4	24	0,750
50	4	32	0,719

Network number 38 (Figure 4.2) composed by 24 LEUs, achieved the maximum overall quality of 1, meaning that, the clusters obtained are all class-uniform. In this case, the network presents only four class labels, that correspond to the number of levels assumed by the target variable turnover class and the number of associated clusters.

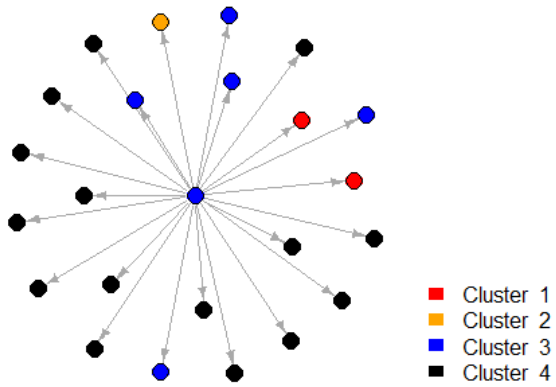


Figure 4.2 - Graphical representation of SUWAN on Network_ID 38, with cluster identification.

From Table 4.4, it can also be retrieved that this network presents the same type of Legal Unit (L), and Limited Liability form (LL) for all observations. The majority of LEUs are from Portugal (PT), with the exception of one, that is settled in Spain (ES).

Furthermore, cluster 1, is composed by two LEUs that belong to class 4 of turnover, that ranges between 10 and 50 million euros. Cluster 2 is classified by turnover class 3, with a range between 2 and 10 million euros. On the other hand, cluster 3 is classified by turnover class of 1, that corresponds to a turnover of zero. This cluster if formed by LEUs with the lowest size class and with the majority of economic activity related to the financial service activities (K.64).

Lastly, cluster 4 is classified by turnover class of 2, categorized by less than 2 million euros of turnover. This could be explained by the fact that most LEUs of this cluster dedicate their economic activity to agriculture and have the lowest size class, corresponding to 0-1 persons employed.

Table 4.4 - Attribute list of nodes from Network_ID 38.

Cluster	Type	Form	Country Code	Size Class	Turnover Class	NACE Div
1	L	LL	ES	1	4	C.10
1	L	LL	PT	4	4	A.01
2	L	LL	PT	4	3	A.01
3	L	LL	PT	1	1	K.64
3	L	LL	PT	1	1	K.64
3	L	LL	PT	1	1	K.64
3	L	LL	PT	1	1	A.01
3	L	LL	PT	1	1	A.01
3	L	LL	PT	1	1	K.64
4	L	LL	PT	1	2	A.01
4	L	LL	PT	1	2	A.01
4	L	LL	PT	1	2	A.01
4	L	LL	PT	1	2	A.01
4	L	LL	PT	1	2	A.01
4	L	LL	PT	1	2	A.01
4	L	LL	PT	1	2	A.01
4	L	LL	PT	1	2	A.01
4	L	LL	PT	1	2	A.01
4	L	LL	PT	1	2	A.01
4	L	LL	PT	1	2	A.01
4	L	LL	PT	1	2	A.01
4	L	LL	PT	1	2	A.01
4	L	LL	PT	2	2	M.69
4	L	LL	PT	1	2	A.01
4	L	LL	PT	1	2	K.64

For network number 25, the SUWAN algorithm grouped the 39 nodes into five clusters, with an overall quality of 0.923. The first cluster has a majority turnover class of 1, with 12 cases. This cluster contains LEUs from type L and form LL. Apart from one case, which operates in Morocco, all LEUs from this cluster operate in Portugal. Also, the size class of the cluster is dispersed around classes 1 and 2, with one observation belonging to class 3. The economic activities of this cluster vary between agriculture and financial service activities.

On the other hand, clusters 2 and 5, with 7 and 14 nodes, respectively, are pure clusters, with turnover class 2. Also, this clusters only contains LEUs with the attributes of type L, form LL, country PT, size class 2 and NACE div A.01.

Cluster 3 just contains a node, that corresponds to the only observation with a turnover class of 4. Finally, cluster 4, with 5 nodes, is also a pure cluster of turnover class 3. This cluster contains LEUs from types L and B, form LL, countries of Portugal and Spain, size class of 4 and 3 and economic activity of agriculture and manufacture of food products.

Network number 48 resulted in 3 clusters, where cluster 1 contains turnover classes of 1 and 2, while clusters 2 and 3 are pure, with turnover classes of 2 and 4, accordingly.

The clustering of network number 51 also revealed the majority of clusters class uniform, except for cluster 4, that contains turnover classes of 1, 5 and 99.

A graphical representation of the networks is followed in Figure 4.3, and the output tables with the observations, attributes and clusters are included in Annex D.

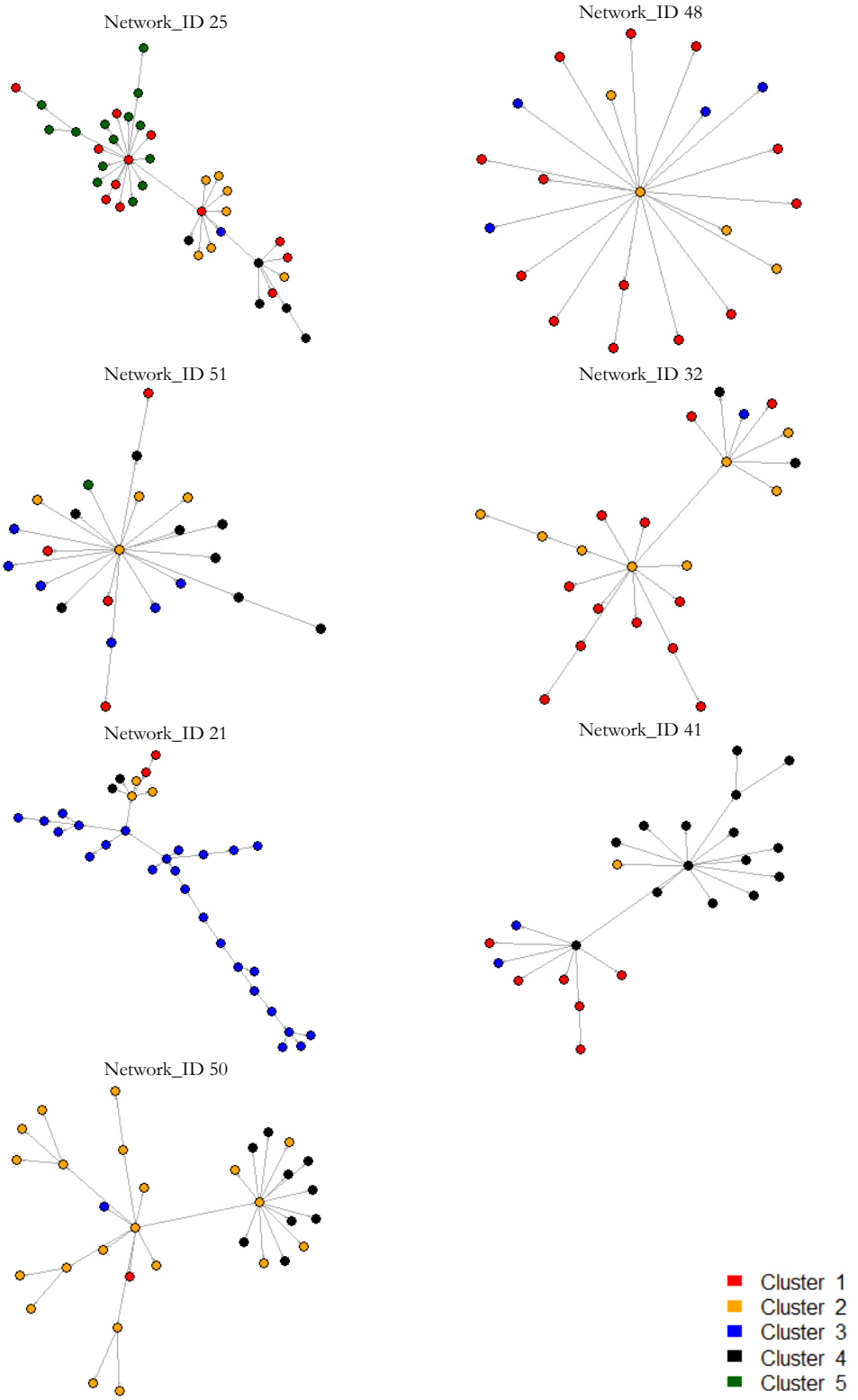


Figure 4.3 - Graphical representation of SUWAN on Networks 25, 48, 51, 32, 21, 41 and 50, with cluster identification.

4.2 Analysis of Subgroup Discovery results

We used Subgoup Discovery (SD) as a benchmark for SUWAN. The results obtained with subgroup discovery method, for the same set of networks, are presented below in Table 4.5. With this algorithm, there are several differences to highlight, like the number of nodes used in each network, that differs from the ones obtained from SUWAN. The reason behind this difference lies in the fact that SD does not group all nodes in clusters, but instead, it finds subgroups of nodes with an associated description. Besides that, the algorithm allows an overlapping of nodes between subgroups, that affects the number of unique nodes used on the SD task.

Table 4.5 - Subgroup discovery results on ERG network, with Portuguese UCI.

Network ID	#Subgroups	#Nodes	#Unique Nodes	Overall Quality
38	4	24	5	0,800
25	4	39	11	0,688
48	3	21	20	0,531
51	3	23	21	0,303
32	4	23	22	0,525
21	4	33	9	0,800
41	4	24	7	0,750
50	1	32	29	0,655

This way, for network number 38, SD generated 4 subgroups, where from the 24 existing nodes, only 5 were grouped. In this specific case, the same five nodes were grouped in four different subgroups, with different descriptions. In fact, the subgroups found are subgroups of each other, with more specific descriptions for the same set of nodes (Table 4.5). The overall quality of this network with subgroup discovery is 80% since there is only one node with a different class label between each subgroup.

From the set of networks presented in Table 4.5, subgroup discovery produced different outcomes from the previous referred. In the case of network number 50, it produced a unique subgroup with 29 observations, based on the description on the economic activity of the LEUs, that correspond to the Human and Health activities. The target class of this subgroup ranges between the values of 2, 3, 4 and 99, with the majority of observations belonging to class 2.

Table 4.6 - Subgroup discovery output for Network_ID 38.

Subgroup	Nodes_ID	Target Class	Description
1	1	1	NACE Div = K.64
	2	2	
	3	1	
	4	1	
	5	1	
2	1	1	NACE Div = K.64 + Country Cod=PT
	2	2	
	3	1	
	4	1	
	5	1	
3	1	1	NACE Div = K.64 + Country Cod=PT + Size Class=1
	2	2	
	3	1	
	4	1	
	5	1	
4	1	1	NACE Div = K.64 + Size Class=1
	2	2	
	3	1	
	4	1	
	5	1	

On the other hand, for network number 48 (Figure 4.3), it grouped 20 of the nodes, with an overlapping of 37.5%. These subgroups are described based on the size class and country code. The first subgroup of size 6, has observations with turnover class of 1 and 2, and it is described by a size class of 2. The second subgroup is a subgroup of the first one, with the addition of the country code (PT) in description. The last subgroup, with 20 nodes, is described by the country code (PT). In this case, the target class varies among the values 1, 2 and 4. Figure 4.4 shows the three subgroups in the same network, separately, due to the overlapping of nodes in the different subgroups. The output tables for subgroup discovery output on networks 25, 48, 51, 32, 21, 41 and 50 are presented in Annex E.

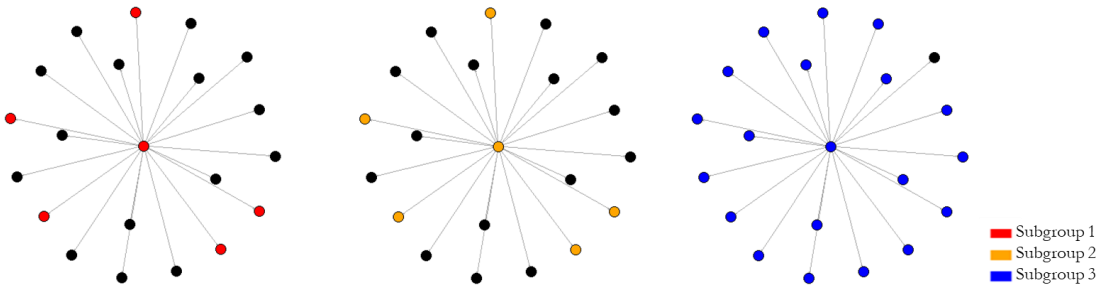


Figure 4.4 - Graphical representation of subgroup discovery on network 48, with colour identification of the subgroups.

Chapter 5

Inter-organizational Performance Analysis

5.1 Variables impact on performance

During the results' analysis for the networks under study it was possible to observe that the variables with higher impact to determine the LEU's turnover class are size class and the economic activity of the group (NACE Div).

For the NACE Div attribute, it is possible to infer that, turnovers of less than 2 million euros (classes 1 to 2) have the most frequent activities of financial services (K.64), real states (L.68) and Professional and scientific and technical activities (M.70). On the other hand, for higher turnover classes of 4 and 5, with more than 10 million euros, the activities with more frequency are in the field of wholesale and retail trade, repair of motor vehicles and motorcycles (G.46, G.45 and G.47), electricity, gas, steam, and air conditioning supply (D.35), manufacturing (C.10, C.16) and construction (F.42 and F.41).

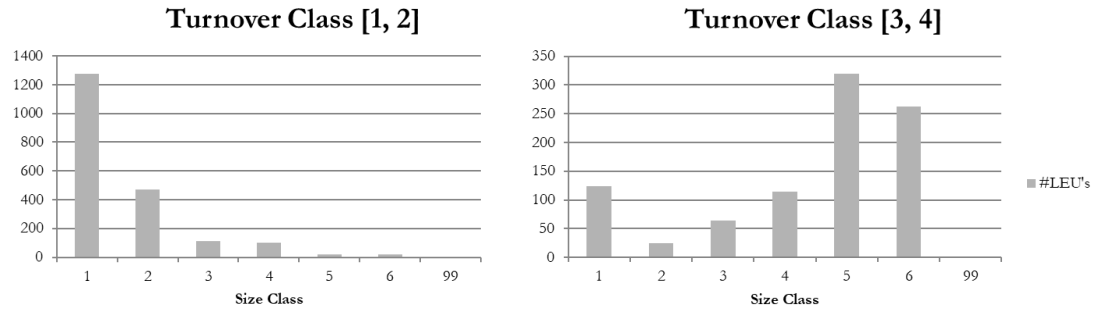


Figure 5.1 - Distribution of LEU's frequencies according to the size class and the turnover class of 1 to 2 and 3 to 4, from ERG network with Portuguese UCI and minimum number of connections of 20.

For the size class, it was possible to observe that the majority of LEU's, with lower turnover classes of 1 and 2, are more condensed in lower size classes of 1 to 2. On the opposite side, LEUs with higher turnover classes have more observations for the size classes of 5 and 6. This dispersion can be observed in Figure 5.1, where the graphs indicate the dispersion of observations of turnover classes from 1 to 2, and 3 to 4, accordingly, with the variation.

Due to the selection of networks with Portuguese UCI, the variables concerning the type and legal form of LEU's proved to be irrelevant for the SUWAN, since that are fewer cases where the attributes type and form are different from L and LL, respectively.

5.1 Network topology impact on performance

In order to analyze the impact of the network's topology on the group's performance, some network topology measures, and the group total turnover were exploited. This way, to study the networks topology, some essential measures can be examined to study the networks compactness, centrality, and density (Table 5.1).

Table 5.1 - Summary on essential network topology measures used to study the impact on the organizational performance.

Measure	Description
Diameter	Measures how compact the network is
Density	Measures the connectivity of the network
Average Degree Centrality	Measures the connectivity of nodes, on average
Average Betweenness Centrality	Measures the capacity of information flow between nodes, on average
Average Closeness Centrality	Measures the influence of nodes in the entire network, on average

The group performance can be measured by its turnover. Hence, a new variable, that indicates the total turnover of the group, was computed, denominated by sum_ent_turnov. Based on the turnover presented by each enterprise that composes the group, the total turnover was obtain through the sum of those values.

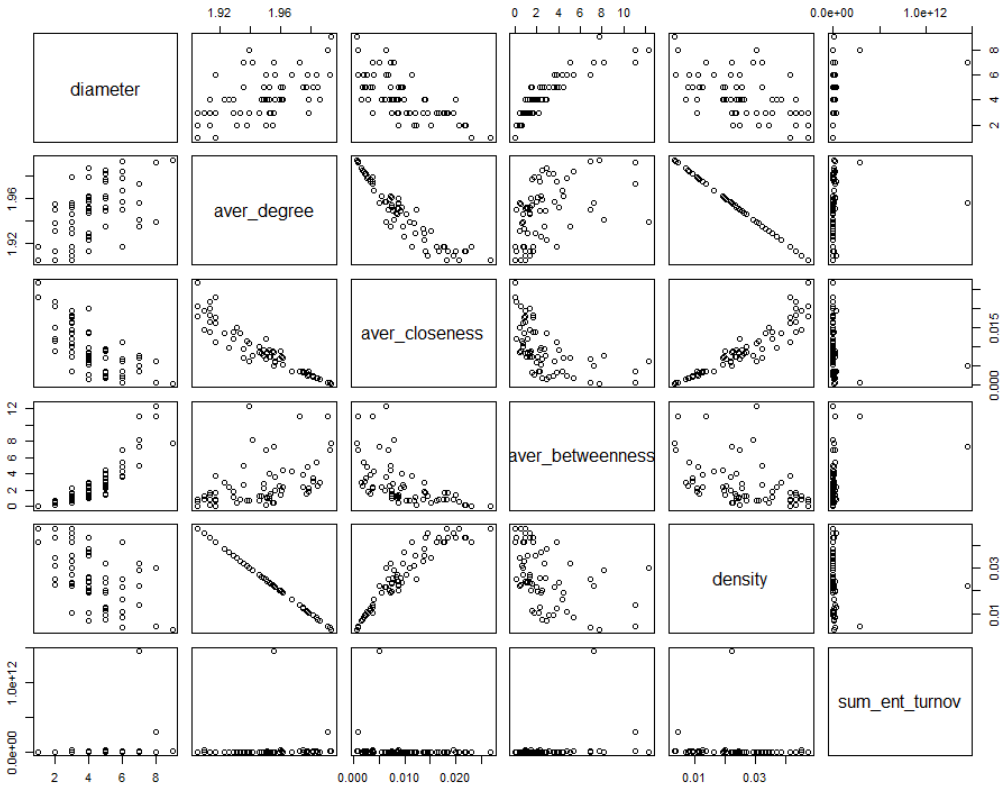


Figure 5.2 - Correlation plot between the variable's diameter, average degree, average closeness, average betweenness, density and total turnover.

Analyzing the correlation between variables in Figure 5.2, it is possible to retain that the pairs of variables average closeness/density and diameter/average betweenness are positively correlated, with correlation values of 0.93 and 0.88, respectively. On the other hand, with a negative correlation value of -0.93 and -0.7, are the pairs of variables average degree/average closeness and diameter/average closeness. Although neither of the variables seems to be correlated with the total turnover of the group, an analysis of a multiple regression model was performed to the 67 networks under analysis. This way, the variables of the total turnover and the network topology measures were placed as dependent and independent variables, accordingly.

Table 5.2 - Multiple linear regression model with network topology measures as independent variables and total turnover as dependent variable.

Residuals	**Min**	**1Q**	**Median**	**3Q**	**Max**
	$-2.126e^{11}$	$-3.546e^{10}$	$-8.559e^{9}$	$8.225e^{9}$	$1.322e^{12}$
Coefficients		**Estimate**	**Std.Error**	**t value**	**Pr(>|t|)**
	(Intercept)	$5.302e^{11}$	$4.814e^{12}$	0.110	0.913
	diameter	$1.188e^{10}$	$3.256e^{10}$	0.365	0.717
	aver_degree	$-3.066e^{11}$	$2.404e^{12}$	-0.128	0.899
	aver_closeness	$7.524e^{11}$	$1.137e^{13}$	0.066	0.947
	aver_betweenness	$1.449e^{10}$	$1.712e^{10}$	0.846	0.401
	density	NA	NA	NA	NA
Residual Std error	$1.767e^{11}$ on 62 degrees of freedom				
Multiple R-squared	0.08537				
Adjusted R-squared	0.02636				
F-statistics	1.447 on 4 and 62 DF, p-value: 0.2294				

When analyzing the multiple linear regression output of Table 5.2, it is possible to observe that the variable that has more impact in the group's turnover is the average closeness, with an estimate value of $7.524e^{11}$, although it does not present a significant p-value. In the same way, none of the variables proved to be significant to the model. Looking at the F-statistics and the overall p-value, it can be concluded that we reject the hypothesis of a relationship between the dependent variables and the independent variables. Therefore, there is no significant evidence that a relationship the total turnover variable and network topology variables exists.

Chapter 6

Conclusions and Challenges

The approach of using both the information about the network structure and the attributes of the nodes in the clustering process proved to be feasible. It enabled the creation of clusters/subgroups that are not only densely linked, but also class-uniform, in terms of the target class that describe those vertices. A characteristic of interest is defined beforehand, denominated by target class, that allows to obtain clusters/subgroups based on a class label.

The application of supervised clustering on attributed networks revealed to be an underdeveloped topic. Atzmueller (2018) applied the subgroup discovery task on attributed social interaction networks. For this, it adapted the principles of subgroup discovery to the dyadic network setting, detecting compositional patterns and capturing subgroup of nodes, estimated by a quality measure. The subgroup discovery was implemented on the EGR networks with the SD-MAP algorithm, using the preprocessing of COMODO algorithm, that combines the graph structure and the descriptive information of the vertices.

On the supervised clustering approach, the SRIDHCR algorithm proposed by Eick (2004), was adapted to consider both structural and compositional characteristics of the EGR network. Moreover, the original algorithm was also adapted for the implementation on categorical variables, through a variation of the k-means, known by k-modes.

In a preliminary analysis of the outputs produced by both methodologies, it was concluded that subgroup discovery produced better clusters/subgroups, with higher overall quality, in comparison with SUWAN. However, subgroup discovery achieved better results due to the lack of nodes grouped, and by allowing an overlapping of nodes between subgroups. On the other hand, the main focus of subgroup discovery is to find subgroups of nodes, described by patterns and with a determined quality measure.

The SUWAN method also produced quite good results, with high-level cluster purity, among the studied cases. This method groups into clusters all nodes of the network, contrary to subgroup discovery.

The focus of the work was to obtain class-uniform clusters, based on the LEUs turnover class, using SUWAN. The analysis of results allowed to verify certain patterns in the nodes that compose the clusters. Clusters with the majority class of turnovers that range between 1 and 2, are formed by LEUs that employ less persons. Similarly, clusters

with class labels of turnover ranging between 5 and 6, are assembled by legal units with size classes of higher levels. Therefore, the turnover is clearly affected by the size of the legal unit.

Additionally, the analysis on the network topology impact on performance proved that there is no significant evidence of a relationship between the total turnover variable and network topology measures of diameter, average degree, closeness and betweenness.

Furthermore, this study revealed that SUWAN in attributed networks involves certain challenges. One of the challenges is the parameterization of variables that influence the clustering output. For example, the importance of the network topology, established by α, has a strong influence and it can provide several different outcomes. Also, the developed methodology works on representative-based supervised clustering, that randomly choses the first k set of representatives. Although this process allows to explore the solution space, the clustering process is still compromised by this randomness. Another challenge is the evaluation method. Evaluating the quality of a clustering is challenging, as the correct clusters are not known. Also, the proposed evaluation method gives more focus on the node's attribute, since it evaluates the overall quality based on the cluster's purity. This may result in circumstances where certain algorithms perform better in terms of network topology but worse in terms of node characteristics, making it difficult to determine which method performs better in the long run.

Bibliography

Al-Harbi, S. H., & Rayward-Smith, V. J. (2006). Adapting k-means for supervised clustering. *Applied Intelligence*, vol. 24, nº3, pp. 219-226. doi:10.1007/s10489-006-8513-8.

Alon, U. (2003). Biological networks: the tinkerer as an engineer. *Science*, vol. 301, nº5641, pp. 1866-1867.

Andrienko, N., Andrienko, G., Savinov, A., Voss, H., & Wettschereck, D. (2001). Exploratory Analysis of Spatial Data Using Interactive Maps and Data Mining. *Cartography and Geographic Information Science*, vol. 28, nº3, pp- 151-166. doi:10.1559/152304001782153035.

Atzmueller, M. (2015). Subgroup discovery. *WIREs Data Mining Knowledge Discovery*, vol. 5, nº1, pp. 35-49. doi:10.1002/widm.1144.

Atzmueller M. (2018). Compositional Subgroup Discovery on Attributed Social Interaction Networks. *In: Soldatova L., Vanschoren J., Papadopoulos G., Ceci M. (eds) Discovery Science. DS 2018. Lecture Notes in Computer Science*, vol. 11198. Springer, Cham. doi:10.1007/978-3-030-01771-2_17.

Atzmueller, M., Doerfel, S., & Mitzlaff, F. (2016). Description-oriented community detection using exhaustive subgroup discovery. *Information Sciences*, vol. 329, 965-984. doi:10.1016/j.ins.2015.05.008.

Atzmueller M., Lemmerich F. (2009). Fast Subgroup Discovery for Continuous Target Concepts. *In: Rauch J., Raś Z.W., Berka P., Elomaa T. (eds) Foundations of Intelligent Systems. ISMIS 2009. Lecture Notes in Computer Science*, vol. 5722. Springer, Berlin, Heidelberg. doi:10.1007/978-3-642-04125-9_7.

Atzmueller, M., & Mitzlaff, F. (2011). Efficient Descriptive Community Mining. *In: Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference*, Palm Beach, Florida, USA.

Atzmueller M., & Puppe F. (2006). SD-Map – A Fast Algorithm for Exhaustive Subgroup Discovery. *In: Fürnkranz J., Scheffer T., Spiliopoulou M. (eds) Knowledge Discovery in Databases: PKDD 2006. PKDD 2006. Lecture Notes in Computer Science*, vol. 4213. Springer, Berlin, Heidelberg. doi:10.1007/11871637_6.

Ball, B., & Newman, M. J. M. (2013). Friendship networks and social status. *Cambridge University Press*, vol. 1, nº1, pp. 16-30.

Carmona, C. J., González, P., Del Jesus, M. J., Romero, C., & Ventura, S. (2010).

Evolutionary algorithms for subgroup discovery applied to e-learning data. *IEEE EDUCON 2010 Conference*, Madrid, pp. 983-990, doi:10.1109/EDUCON.2010.5492470.

Chavent, M., Kuentz-Simonet, V., Labenne, A., & Saracco, J. (2018). ClustGeo: an R package for hierarchical clustering with spatial constraints. *Computational Statistics*, vol. 33, nº4, pp. 1799-1822, doi:10.1007/s00180-018-0791-1.

Combe, D., Largeron, C., Géry, M., & Egyed-Zsigmond, E. (2015). I-louvain: An attributed graph clustering method. In Fromont, E., De Bie, T., and van Leeuwen, M., editors, *Advances in Intelligent Data Analysis XIV*, pp. 181–192, Cham. Springer International Publishing.

Dang, T., & Viennet, E. (2012). Community detection based on structural and attribute similarities. In *Proceedings of the Sixth International Conference on Digital Society*, pp. 7–12.

Deng, J., Kang, B., Lijffijt, J., & Bie, T. (2020). Explainable Subgraphs with Surprising Densities: A Subgroup Discovery Approach. ArXiv abs/2002.00793.

Ebers, M. (1999). The formation of inter-organizational networks. *Oxford University Press*.

Eick, C. F., & Zeidat, N. (2004). K-mendoid-style Clustering Algorithms for Supervised Summary Generation. *In: Proceedings of the International Conference on Artificial Intelligence, 2004*.

Eurostat (2010). Business Registers Recommendations Manual, *Methodologies and Working papers*, Publication Office of the European Union, Luxembourg.

Falih, I., Grozavu, N., Kanawati, R., & Bennani, Y. (2018). ANCA: Attributed Network Clustering Algorithm. In Cherifi, C., Cherifi, H., Karsai, M., and Musolesi, M., editors, *Complex Networks & Their Applications VI*, pp. 241–252, Cham. Springer International Publishing.

Finley, T., & Joachims, T. (2008). Supervised k-Means Clustering.

Fortunato, S., & Hric, D. (2016). Community detection in networks: A user guide. *Physics Reports*, vol. 659, pp. 1-44. doi:10.1016/j.physrep.2016.09.002.

Freeman, L. (2004). The Development of Social Network Analysis. *Empirical Press*.

Freeman, L. (1996). Cliques, Galois lattices, and the structure of human social groups. *Social networks*, vol. 18, nº3, pp. 173-187.

Gamberger, D., & Lavrac, N. (2002). Generating Actionable Knowledge by Expert-Guided Subgroup Discovery. *In: Elomaa T., Mannila H., Toivonen H. (eds) Principles of Data Mining and Knowledge Discovery. PKDD 2002. Lecture Notes in Computer Science (Lecture*

Notes in Artificial Intelligence), vol 2431. Springer, Berlin, Heidelberg. doi:10.1007/3-540-45681-3_14.

Gan, H., Huang, R., Luo, Z., Xi, X., & Gao, Y. (2018). On using supervised clustering analysis to improve classification performance. *Information Sciences*, vols. 454-455, pp. 216-228. doi:10.1016/j.ins.2018.04.080.

Greiner, B., & Levati, M. V. (2005). Indirect reciprocity in cyclical networks: An experimental study. *Journal of Economic Psychology*, vol. 26, nº5, pp. 711-731.

Harenberg, S., Bello, G., Gjeltema, L., Ranshous, S., Harlalka, J., Seay, R., Padmanabhan, K., & Samatova, N. (2014). Community detection in large-scale networks: a survey and empirical evaluation. *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 6, nº6, pp. 426-439. doi:10.1002/wics.1319.

Harvey, N. J. A., Kleinberg, R., & Lehman, A. R. (2006). On the capacity of information networks. *IEEE Transactions on Information Theory*, vol. 52, nº6, pp. 2345-2364. doi:10.1109/TIT.2006.874531.

Helal, S. (2016). Subgroup Discovery Algorithms: A Survey and Empirical Evaluation. *Journal of Computer Science and Technology*, vol. 31, pp. 561–576. doi:10.1007/s11390-016-1647-1.

Herrera, F., Carmona, C. J., González, P., & Del Jesus, M. J. (2011). An overview on subgroup discovery: foundations and applications. *Knowledge Information Systems*, vol. 29, nº3, pp. 495-525.

Hewapathirana, I. U. (2019). Change detection in dynamic attributed networks. *WIREs Data Mining and Knowledge Discovery*, vol. 9, nº3. doi:10.1002/widm.1286.

Hoberecht, S., Joseph, B., Spencer, J., & Southern, N. (2011). Inter-organizational networks: An emerging paradigm of whole systems change. *Journal of the Organization Development Network*, vol. 43, nº4, pp. 23-27.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, vol. 31, nº8, pp. 651-666. doi:10.1016/j.patrec.2009.09.011.

Kavšek, B., & Lavrac, N. (2006). APRIORI-SD: Adapting association rule learning to subgroup discovery. *Applied Artificial Intelligence*, vol. 20, nº7, pp. 230-241.

Klösgen, W. (1996). Explora: A multipattern and multistrategy discovery assistant. *Advances in Knowledge Discovery and Data Mining*, pp. 249-271.

Klösgen, W., & May, M. (2002). Census data mining—an application. *In: Proceedings of the 6th European conference on principles of data mining and knowledge discovery*, pp. 65–79.

Lavrač, N., Kavšek, B., Flach, P., & Todorovski, L. (2004). Subgroup discovery with CN2-SD. *Journal of Machine Learning Research*, vol. 5, pp. 153-188.

Lechner, C., & Dowling, M. (2003), "Firm Networks: external relationships as sources for the growth and competitiveness of entrepreneurial firms", Entrepreneurship & Regional Development, vol. 15, pp.1-26.

Lucas, T., Gomes, J., Vimieiro, R., Prudêncio, R., & Ludermir, T. (2019). A Multivariate Method for Group Profiling Using Subgroup Discovery. *In: 2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, Salvador, Brazil, pp. 371-376, doi:10.1109/BRACIS.2019.00072.

Matous, P., & Todo, Y. (2017). Analyzing the coevolution of interorganizational networks and organizational performance: Automakers' production networks in Japan. *Applied Network Science*, vol. 2, nº1, pp. 5. doi:10.1007/s41109-017-0024-5.

Meeng, M., & Knobbe, A. (2020). For real: a thorough look at numeric attributes in subgroup discovery. *Data Mining and Knowledge Discovery*, vol. 35 pp. 1-55. doi:10.1007/s10618-020-00703.

Mueller, M., Rosales, R., Steck, H., Krishnan, S., Rao, B., & Kramer, S. (2009). Subgroup Discovery for Test Selection: A Novel Approach and Its Application to Breast Cancer Diagnosis. *In: Adams N.M., Robardet C., Siebes A., Boulicaut JF. (eds) Advances in Intelligent Data Analysis VIII. IDA 2009. Lecture Notes in Computer Science*, vol. 5772. Springer, Berlin, Heidelberg. doi:10.1007/978-3-642-03915-7_11.

Newman, M. E. J. (2006). Modularity and community structure in networks. *In: Proceedings of the National Academy of Sciences*, vol. 103, nº 23, pp. 8577-8582. doi:10.1073/pnas.0601602103.

Oliveira, M., & Gama, J. (2012). An overview of social network analysis. *WIREs Data Mining and Knowledge Discovery*, vol. 2, pp. 99-115. doi:10.1002/widm.1048.

Palla, G., Farkas, I. J., Pollner, P., Derényi, I., & Vicsek, T. (2007). Directed network modules. *New Journal of Physics*, vol. 9, nº 6, pp. 186.

Pandit, S., & Gupta, S. (2011). A comparative study on distance measuring approaches for clustering. *International Journal of Research in Computer Science*, vol. 2, nº1, pp. 29-31.

Popp, J., Milward, H. B., MacKean, G., Casebeer, A., & Lindstrom, R. (2014). Inter-organizational networks: A review of the literature to inform practice. *IBM Center for the Business of Government*, pp. 93-96.

Ricciardi, F., & Rossignoli, C. (2015). Inter-Organizational Relationships. Towards a

Dynamic Model for Understanding Business Network Performance. *Springer International Publishing*.

Singh, H., & Sharma, R. (2012). Role of adjacency matrix & adjacency list in graph theory. *International Journal of Computers & Technology*, vol. 3, nº 1, pp. 179-183.

Tabassum, S., & Pereira, F. S., & Fernandes, S., & Gama, J. (2018). Social network analysis: An overview. WIREs Data Mining and Knowledge Discovery, vol. 8, nº 5, e1256. doi:10.1002/widm.1256.

Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications. *Cambridge University Press*.

Wrobel S. (1997). An algorithm for multi-relational discovery of subgroups. *In: Komorowski J., Zytkow J. (eds) Principles of Data Mining and Knowledge Discovery. PKDD 1997. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence)*, vol. 1263. Springer, Berlin, Heidelberg. doi:10.1007/3-540-63223-9_108.

Yang, W.-S., Dia, J.-B., Cheng, H.-C., & Lin, H.-T. (2006). Mining social networks for targeted advertising. *In: Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, Kauia, HI, USA, 2006, pp. 137a-137a, doi:10.1109/HICSS.2006.272.

Annex A

Attributes description for the EGR data base

Table A.1 - Description of types of LEUs.

LEU_TYPE	Description
B	Branch
L	Legal unit (not branch)

Table A.2 - Description of forms of LEUs.

LEU_LFORM	Description
LL	Limited liability company - include limited liability partnerships and public corporations
SP	Sole proprietor
PA	Partnership - exclude limited liability partnerships
GO	Government - local and central government - exclude public corporations
NB	Nonprofit body or mutual association
NP	Natural person - include only if not involved in any economic activity
ND	Not defined - units should be coded here only temporarily

Table A.3 - Description of 2-digit ISO country codes.

COUNTRY_CODE	Country Name
AD	Andorra
AE	United Arab Emirates
AF	Afghanistan
AG	Antigua And Barbuda
AI	Anguilla
AL	Albania
AM	Armenia
AO	Angola
AQ	Antarctica
AR	Argentina
AS	American Samoa
AT	Austria
AU	Australia
AW	Aruba
AX	Aland Islands
AZ	Azerbaijan
BA	Bosnia And Herzegovina
BB	Barbados
BD	Bangladesh
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BH	Bahrain
BI	Burundi
BJ	Benin
BL	Saint Barthelemy
BM	Bermuda
BN	Brunei Darussalam
BO	Bolivia
BR	Brazil
BS	Bahamas

COUNTRY_CODE	Country Name
BT	Bhutan
BV	Bouvet Island
BW	Botswana
BY	Belarus
BZ	Belize
CA	Canada
CC	Cocos (Keeling) Islands
CD	Congo, The Democratic Republic Of
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte D'ivoire
CK	Cook Islands
CL	Chile
CM	Cameroon
CN	China
CO	Colombia
CR	Costa Rica
CU	Cuba
CV	Cape Verde
CX	Christmas Island
CY	Cyprus
CZ	Czech Republic
DE	Germany
DJ	Djibouti
DK	Denmark
DM	Dominica
DO	Dominican Republic
DZ	Algeria
EC	Ecuador
EE	Estonia
EG	Egypt
EH	Western Sahara
ER	Eritrea
ES	Spain
ET	Ethiopia
FI	Finland
FJ	Fiji
FK	Falkland Islands (Malvinas)
FM	Micronesia, Federated States Of
FO	Faroe Islands
FR	France
GA	Gabon
GB	United Kingdom
GD	Grenada
GE	Georgia
GF	French Guiana
GG	Guernsey
GH	Ghana
GI	Gibraltar
GL	Greenland
GM	Gambia
GN	Guinea
GP	Guadeloupe
GQ	Equatorial Guinea
GR	Greece
GS	South Georgia And The South Sandwich Islands
GT	Guatemala
GU	Guam
GW	Guinea-Bissau
GY	Guyana

COUNTRY_CODE	Country Name
HK	Hong Kong
HM	Heard Island And Mcdonald Islands
HN	Honduras
HR	Croatia
HT	Haiti
HU	Hungary
ID	Indonesia
IE	Ireland
IL	Israel
IM	Isle Of Man
IN	India
IO	British Indian Ocean Territory
IQ	Iraq
IR	Iran, Islamic Republic Of
IS	Iceland
IT	Italy
JE	Jersey
JM	Jamaica
JO	Jordan
JP	Japan
KE	Kenya
KG	Kyrgyzstan
KH	Cambodia
KI	Kiribati
KM	Comoros
KN	Saint Kitts And Nevis
KP	Korea, Democratic People's Republic Of
KR	Korea, Republic Of
KW	Kuwait
KY	Cayman Islands
KZ	Kazakhstan
LA	Lao People's Democratic Republic
LB	Lebanon
LC	Saint Lucia
LI	Liechtenstein
LK	Sri Lanka
LR	Liberia
LS	Lesotho
LT	Lithuania
LU	Luxembourg
LV	Latvia
LY	Libyan Arab Jamahiriya
MA	Morocco
MC	Monaco
MD	Moldova, Republic Of
ME	Montenegro
MF	Saint Martin (French Part)
MG	Madagascar
MH	Marshall Islands
MK	The Republic Of North Macedonia
ML	Mali
MM	Myanmar
MN	Mongolia
MO	Macao
MP	Northern Mariana Islands
MQ	Martinique
MR	Mauritania
MS	Montserrat
MT	Malta
MU	Mauritius
MV	Maldives

COUNTRY_CODE	Country Name
MW	Malawi
MX	Mexico
MY	Malaysia
MZ	Mozambique
NA	Namibia
NC	New Caledonia
NE	Niger
NF	Norfolk Island
NG	Nigeria
NI	Nicaragua
NL	Netherlands
NO	Norway
NP	Nepal
NR	Nauru
NU	Niue
NZ	New Zealand
OM	Oman
PA	Panama
PE	Peru
PF	French Polynesia
PG	Papua New Guinea
PH	Philippines
PK	Pakistan
PL	Poland
PM	Saint Pierre And Miquelon
PN	Pitcairn
PR	Puerto Rico
PS	Palestine, State Of
PT	Portugal
PW	Palau
PY	Paraguay
QA	Qatar
RE	Reunion
RO	Romania
RS	Serbia
RU	Russian Federation
RW	Rwanda
SA	Saudi Arabia
SB	Solomon Islands
SC	Seychelles
SD	Sudan
SE	Sweden
SG	Singapore
SH	Saint Helena
SI	Slovenia
SJ	Svalbard And Jan Mayen
SK	Slovakia
SL	Sierra Leone
SM	San Marino
SN	Senegal
SO	Somalia
SR	Suriname
ST	Sao Tome And Principe
SV	El Salvador
SY	Syrian Arab Republic
SZ	Swaziland
TC	Turks And Caicos Islands
TD	Chad
TF	French Southern Territories
TG	Togo
TH	Thailand

COUNTRY_CODE	Country Name
TJ	Tajikistan
TK	Tokelau
TL	Timor-Leste
TM	Turkmenistan
TN	Tunisia
TO	Tonga
TR	Turkey
TT	Trinidad And Tobago
TV	Tuvalu
TW	Taiwan, Province Of China
TZ	Tanzania, United Republic Of
UA	Ukraine
UG	Uganda
UM	United States Minor Outlying Islands
US	United States
UY	Uruguay
UZ	Uzbekistan
VA	Holy See (Vatican City State)
VC	Saint Vincent And The Grenadines
VE	Venezuela
VG	Virgin Islands, British
VI	Virgin Islands, U.S.
VN	Viet Nam
VU	Vanuatu
WF	Wallis And Futuna
WS	Samoa
YE	Yemen
YT	Mayotte
ZA	South Africa
ZM	Zambia
ZW	Zimbabwe
ZZ	Neutral Zone
II	Supranational
BQ	Bonaire, Sint Eustatius And Saba
CW	Curacao
SX	Sint Maarten (Dutch Part)
SS	South Sudan

Table A.4 - Size of the enterprise based on number of persons employed.

SIZE_CLASS	Description
1	0-1
2	2-9 persons employed
3	10-19 persons employed
4	20-49 persons employed
5	50-249 persons employed
6	250 or more persons employed
99	Not defined

Table A.5 - Turnover class based on the enterprise turnover values.

TURNOVER_CLASS	Description
1	0
2	Less than 2 million
3	Between 2 and 10 million
4	Between 10 and 50 million
5	More than 50 million
99	Not defined

Table A.6 - NACE Rev. 2 activity codes for the main activity of enterprises, with junction of section and division code.

NACE SEC+DIV CODE	Description
A.01	Crop and animal production, hunting and related service activities
A.02	Forestry and logging
A.03	Fishing and aquaculture
B.05	Mining of coal and lignite
B.06	Extraction of crude petroleum and natural gas
B.07	Mining of metal ores
B.08	Other mining and quarrying
B.09	Mining support service activities
C.10	Manufacture of food products
C.11	Manufacture of beverages
C.12	Manufacture of tobacco products
C.13	Manufacture of textiles
C.14	Manufacture of wearing apparel
C.15	Manufacture of leather and related products
C.16	Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials
C.17	Manufacture of paper and paper products
C.18	Printing and reproduction of recorded media
C.19	Manufacture of coke and refined petroleum products
C.20	Manufacture of chemicals and chemical products
C.21	Manufacture of basic pharmaceutical products and pharmaceutical preparations
C.22	Manufacture of rubber and plastic products
C.23	Manufacture of other non-metallic mineral products
C.24	Manufacture of basic metals
C.25	Manufacture of fabricated metal products, except machinery and equipment
C.26	Manufacture of computer, electronic and optical products
C.27	Manufacture of electrical equipment
C.28	Manufacture of machinery and equipment n.e.c.
C.29	Manufacture of motor vehicles, trailers and semi-trailers
C.30	Manufacture of other transport equipment
C.31	Manufacture of furniture
C.32	Other manufacturing
C.33	Repair and installation of machinery and equipment
D.35	Electricity, gas, steam and air conditioning supply
E.36	Water collection, treatment and supply
E.37	Sewerage
E.38	Waste collection, treatment and disposal activities; materials recovery
E.39	Remediation activities and other waste management services
F.41	Construction of buildings
F.42	Civil engineering
F.43	Specialized construction activities
G.45	Wholesale and retail trade and repair of motor vehicles and motorcycles
G.46	Wholesale trade, except of motor vehicles and motorcycles
G.47	Retail trade, except of motor vehicles and motorcycles
H.49	Land transport and transport via pipelines
H.50	Water transport
H.51	Air transport
H.52	Warehousing and support activities for transportation
H.53	Postal and courier activities
I.55	Accommodation
I.56	Food and beverage service activities
J.58	Publishing activities
J.59	Motion picture, video and television program production, sound recording and music publishing activities
J.60	Programming and broadcasting activities
J.61	Telecommunications
J.62	Computer programming, consultancy and related activities

NACE SEC+DIV CODE	Description
J.63	Information service activities
K.64	Financial service activities, except insurance and pension funding
K.65	Insurance, reinsurance and pension funding, except compulsory social security
K.66	Activities auxiliary to financial services and insurance activities
L.68	Real estate activities
M.69	Legal and accounting activities
M.70	Activities of head offices; management consultancy activities
M.71	Architectural and engineering activities; technical testing and analysis
M.72	Scientific research and development
M.73	Advertising and market research
M.74	Other professional, scientific and technical activities
M.75	Veterinary activities
N.77	Rental and leasing activities
N.78	Employment activities
N.79	Travel agency, tour operator reservation service and related activities
N.80	Security and investigation activities
N.81	Services to buildings and landscape activities
N.82	Office administrative, office support and other business support activities
O.84	Public administration and defense; compulsory social security
P.85	Education
Q.86	Human health activities
Q.87	Residential care activities
Q.88	Social work activities without accommodation
R.90	Creative, arts and entertainment activities
R.91	Libraries, archives, museums and other cultural activities
R.92	Gambling and betting activities
R.93	Sports activities and amusement and recreation activities
S.94	Activities of membership organizations
S.95	Repair of computers and personal and household goods
S.96	Other personal service activities
T.97	Activities of households as employers of domestic personnel
T.98	Undifferentiated goods-and services-producing activities of private households for own use
U.99	Activities of extraterritorial organizations and bodies

Annex B

EGR Networks of 2018 under analysis

Table B.1 - Networks of 2018, with Portuguese UCI and minimum number of nodes of 20.

Network_ID	GEG_N_LEU	GEG_N_ENT
1	31	31
2	37	37
3	43	43
4	41	37
5	24	23
6	61	59
7	42	36
8	45	35
9	21	21
10	52	51
11	26	25
12	23	21
13	30	28
14	33	24
15	108	99
16	92	82
17	73	69
18	34	29
19	38	37
20	22	22
21	33	29
22	23	21
23	21	21
24	109	96
25	39	37
26	27	26
27	28	27
28	44	41
29	40	36
30	80	77
31	22	21
32	23	23
33	95	35
34	42	39
35	31	29
36	46	43
37	50	48
38	24	24
39	24	23
40	24	22
41	24	24
42	95	91
43	40	37
44	88	83
45	146	71
46	42	41
47	41	32
48	21	21
49	23	22
50	32	31
51	23	23
52	37	36

Network_ID	GEG_N_LEU	GEG_N_ENT
53	28	28
54	29	27
55	46	40
56	79	63
57	133	98
58	52	52
59	49	49
60	45	42
61	51	48
62	23	21
63	39	39
64	327	265
65	276	187
66	235	217
67	123	119

Annex C

Comparison of performance between methods

Table C.1 - Quality results of SUWAN and subgroup discovery on ERG networks with Portuguese UCI.

		SUWAN		Subgroup Discovery		
Network ID	**#Nodes**	**#Clusters**	**Overall Quality**	**#Subgroups**	**#Unique Nodes**	**Overall Quality**
1	31	2	0,581	4	19	0,939
2	37	4	0,541	3	16	0,591
3	43	1	0,535	4	25	0,808
4	41	5	0,610	4	16	0,700
5	24	4	0,500	4	21	0,487
6	61	3	0,377	4	4	0,500
7	42	4	0,500	4	38	0,595
8	45	2	0,378	4	5	1,000
9	21	4	0,524	4	18	0,500
10	52	4	0,385	4	45	0,392
11	26	5	0,654	4	3	1,000
12	23	4	0,696	3	3	1,000
13	30	2	0,367	4	9	0,750
14	33	5	0,545	4	5	1,000
15	108	3	0,417	4	97	0,262
16	92	5	0,478	4	4	1,000
17	73	2	0,329	3	14	0,545
18	34	3	0,412	4	9	0,800
19	38	2	0,421	4	2	1,000
20	22	2	0,364	1	2	1,000
21	33	4	0,758	4	9	0,800
22	23	1	0,304	4	18	0,542
23	21	1	0,619	4	3	1,000
24	109	3	0,404	4	96	0,265
25	39	5	0,923	4	11	0,688
26	27	2	0,444	4	4	1,000
27	28	2	0,500	4	12	0,950
28	44	1	0,523	4	31	0,625
29	40	3	0,375	4	9	0,750
30	80	3	0,688	4	61	0,538
31	22	2	0,409	4	8	0,682
32	23	4	0,783	4	22	0,525
33	95	2	0,558	4	50	1,000
34	42	3	0,643	4	33	0,857
35	31	2	0,548	4	28	0,625
36	46	3	0,348	4	12	0,833
37	50	4	0,620	4	43	0,415
38	24	4	1,000	4	5	0,800
39	24	4	0,458	3	20	0,423
40	24	3	0,375	4	2	1,000
41	24	4	0,750	4	7	0,750
42	95	2	0,305	4	2	1,000
43	40	1	0,375	1	36	0,417
44	88	5	0,432	4	37	0,627
45	146	6	0,651	4	70	0,869

		SUWAN		Subgroup Discovery		
Network ID	**#Nodes**	**#Clusters**	**Overall Quality**	**#Subgroups**	**#Unique Nodes**	**Overall Quality**
46	42	3	0,524	4	19	0,789
47	41	4	0,683	4	3	1,000
48	21	3	0,905	3	20	0,531
49	23	2	0,261	4	8	0,938
50	32	4	0,719	1	29	0,655
51	23	5	0,870	3	21	0,303
52	37	2	0,270	3	34	0,303
53	28	3	0,571	4	16	0,500
54	29	4	0,552	4	23	0,586
55	46	5	0,500	4	6	1,000
56	79	5	0,405	4	12	1,000
57	133	6	0,489	4	7	1,000
58	52	3	0,615	4	3	1,000
59	49	2	0,592	4	25	0,804
60	45	4	0,667	4	39	0,321
61	51	4	0,529	4	2	1,000
62	23	3	0,609	4	12	0,783
63	39	2	0,359	4	17	0,444
64	327	3	0,514	4	21	1,000
65	276	5	0,536	4	72	0,389
66	235	3	0,455	4	8	1,000
67	123	6	0,593	4	81	0,416

Annex D

SUWAN output tables

Table D.1 - List of attributes for the observations of network number 25, with cluster identification.

Cluster	Type	Form	Country Code	Size Class	Turnover Class	NACE Div
1	L	99	MA	1	99	99
1	L	LL	PT	3	3	A.01
1	L	LL	PT	2	1	A.01
1	L	LL	PT	1	1	A.01
1	L	LL	PT	1	1	K.64
1	L	LL	PT	1	1	A.01
1	L	LL	PT	1	1	A.01
1	L	LL	PT	2	3	A.01
1	L	LL	PT	1	1	K.64
1	L	LL	PT	1	1	K.64
1	L	LL	PT	1	1	A.01
1	L	LL	PT	1	1	A.01
2	L	LL	PT	2	2	A.01
2	L	LL	PT	2	2	A.01
2	L	LL	PT	2	2	A.01
2	L	LL	PT	2	2	A.01
2	L	LL	PT	2	2	A.01
2	L	LL	PT	2	2	A.01
2	L	LL	PT	2	2	A.01
3	L	LL	PT	2	4	C.10
4	L	LL	ES	4	3	A.01
4	L	LL	ES	4	3	A.01
4	B	LL	ES	4	3	C.10
4	L	LL	PT	3	3	A.01
4	L	LL	PT	4	3	A.01
5	L	LL	PT	2	2	A.01
5	L	LL	PT	2	2	A.01
5	L	LL	PT	2	2	A.01
5	L	LL	PT	2	2	A.01
5	L	LL	PT	2	2	A.01
5	L	LL	PT	2	2	A.01
5	L	LL	PT	2	2	A.01
5	L	LL	PT	2	2	A.01
5	L	LL	PT	2	2	A.01
5	L	LL	PT	2	2	A.01
5	L	LL	PT	2	2	A.01
5	L	LL	PT	2	2	A.01
5	L	LL	PT	2	2	A.01
5	L	LL	PT	2	2	A.01

Table D.2 - List of attributes for the observations of network number 48, with cluster identification.

Cluster	Type	Form	Country Code	Size Class	Turnover Class	NACE Div
1	L	LL	PT	1	1	R.93
1	L	LL	PT	1	1	I.56
1	L	LL	PT	2	2	R.93
1	L	LL	PT	1	1	I.55
1	L	LL	PT	2	1	H.50
1	L	LL	PT	3	2	R.93
1	L	LL	PT	1	1	I.55
1	L	LL	PT	2	1	R.93
1	L	LL	PT	1	1	M.70
1	L	LL	PT	1	1	K.64
1	L	LL	PT	2	1	I.55
1	L	LL	PT	1	1	R.93
1	L	LL	PT	1	1	R.93
2	L	LL	PT	1	2	H.52
2	L	LL	PT	2	2	K.64
2	L	LL	PT	2	2	L.68
2	L	LL	PT	5	2	I.55
3	L	LL	DE	5	4	N.79
3	L	LL	PT	6	4	R.93
3	L	LL	PT	5	4	R.93
3	L	LL	PT	5	4	R.93

Table D.3 - List of attributes for the observations of network number 51, with cluster identification.

Cluster	Type	Form	Country Code	Size Class	Turnover Class	NACE Div
1	L	LL	FR	3	3	G.46
1	L	LL	PT	4	3	C.10
1	L	LL	PT	1	3	A.01
1	L	LL	PT	2	3	A.01
2	L	LL	PT	3	2	A.01
2	L	LL	PT	1	2	D.35
2	L	LL	PT	1	2	M.70
2	L	LL	PT	3	2	I.56
3	L	LL	PT	5	4	A.01
3	L	LL	PT	6	4	H.49
3	L	LL	PT	1	4	A.01
3	L	LL	PT	4	4	G.46
3	L	LL	PT	5	4	A.01
3	L	LL	PT	5	4	G.47
4	L	LL	LU	1	99	G.46
4	L	LL	PT	5	5	C.10
4	L	LL	PT	6	5	C.10
4	L	LL	PT	6	5	C.10
4	L	LL	PT	6	5	G.47
4	L	LL	PT	5	5	C.10
4	L	LL	PT	1	1	C.10
4	L	LL	PT	1	1	C.10
5	L	LL	PT	2	2	G.45

Table D.4 - List of attributes for the observations of network number 32, with cluster identification.

Cluster	Type	Form	Country Code	Size Class	Turnover Class	NACE Div
1	99	99	AO	1	99	99
1	L	LL	LU	1	99	K.64
1	L	LL	PT	1	2	L.68
1	L	LL	PT	2	3	M.70
1	L	LL	PT	1	2	M.70
1	L	LL	PT	3	3	G.47
1	L	LL	PT	2	2	L.68
1	L	LL	PT	1	1	L.68
1	L	LL	PT	2	2	N.82
1	L	LL	PT	1	2	M.70
1	L	LL	PT	1	2	M.70
1	L	LL	PT	1	2	L.68
2	L	LL	PT	2	1	C.23
2	L	LL	PT	1	1	K.64
2	L	LL	PT	1	1	G.47
2	L	LL	PT	1	1	M.70
2	L	LL	PT	4	1	M.70
2	L	LL	PT	1	1	K.64
2	L	LL	PT	1	1	H.52
2	L	LL	PT	1	1	B.08
3	L	LL	PT	1	2	G.46
4	L	LL	PT	5	4	G.46
4	L	LL	PT	6	4	G.47

Table D.5 - List of attributes for the observations of network number 21, with cluster identification.

Cluster	Type	Form	Country Code	Size Class	Turnover Class	NACE Div
1	L	LL	PT	1	2	A.01
1	L	LL	PT	1	2	A.01
2	L	LL	ES	6	5	C.10
2	L	LL	PT	5	5	C.10
2	L	LL	US	5	5	G.46
3	L	PA	BR	2	5	G.46
3	L	LL	ES	6	5	C.10
3	L	LL	ES	6	5	C.10
3	L	LL	ES	6	5	C.10
3	L	LL	99	99	99	99
3	L	LL	PT	1	1	K.64
3	L	LL	PT	1	1	C.10
3	L	LL	PT	5	5	C.10
3	L	LL	PT	3	2	M.70
3	L	LL	PT	1	1	L.68
3	L	LL	PT	1	1	M.70
3	L	LL	PT	1	1	K.64
3	L	LL	PT	1	1	K.64
3	L	LL	PT	1	1	K.64
3	L	LL	PT	1	1	K.64
3	L	LL	PT	2	1	K.64
3	L	LL	PT	1	1	M.70
3	L	LL	PT	2	2	L.68
3	L	LL	PT	1	1	F.41
3	L	LL	PT	1	1	K.64
3	L	LL	PT	1	1	L.68
3	L	LL	PT	1	1	K.64
3	L	LL	PT	1	1	N.77
3	L	LL	PT	1	1	K.64
3	L	LL	PT	1	1	D.35
3	L	LL	PT	1	1	C.10
4	L	LL	PT	1	4	C.10
4	L	ND	TN	1	4	G.46

Table D.6 - List of attributes for the observations of network number 41, with cluster identification.

Cluster	Type	Form	Country Code	Size Class	Turnover Class	NACE Div
1	L	LL	PT	1	2	A.01
1	L	LL	PT	2	2	G.46
1	L	LL	PT	2	2	L.68
1	L	LL	PT	1	2	A.01
1	L	LL	PT	1	2	F.41
1	L	LL	PT	2	2	G.47
2	L	LL	PT	3	3	G.47
3	L	LL	PT	2	3	H.49
3	L	LL	PT	3	3	G.46
4	L	LL	ES	1	2	G.46
4	L	LL	PT	5	4	G.47
4	L	LL	PT	4	4	G.46
4	L	LL	PT	1	2	G.47
4	L	LL	PT	2	2	G.47
4	L	LL	PT	2	1	K.64
4	L	LL	PT	1	2	G.47
4	L	LL	PT	1	1	K.64
4	L	LL	PT	1	2	L.68
4	L	LL	PT	2	3	G.46
4	L	LL	PT	1	1	M.73
4	L	LL	PT	2	2	M.75
4	L	LL	PT	4	2	M.69
4	L	LL	PT	2	2	M.70
4	L	LL	PT	3	2	J.62

Table D.7 - List of attributes for the observations of network number 50, with cluster identification.

Cluster	Type	Form	Country Code	Size Class	Turnover Class	NACE Div
1	L	LL	PT	5	3	Q.86
2	L	LL	#N/D	99	99	Q.86
2	L	LL	PT	5	4	Q.86
2	L	LL	PT	6	4	Q.86
2	L	LL	PT	4	3	Q.86
2	L	LL	PT	5	3	Q.86
2	L	LL	PT	2	2	Q.86
2	L	LL	PT	2	2	Q.86
2	L	LL	PT	5	3	Q.86
2	L	LL	PT	2	2	Q.86
2	L	LL	PT	4	3	Q.86
2	L	LL	PT	4	2	Q.86
2	L	LL	PT	2	2	Q.86
2	L	LL	PT	2	2	Q.86
2	L	LL	PT	4	3	Q.86
2	L	LL	PT	2	2	Q.86
2	L	LL	PT	5	3	Q.86
2	L	LL	PT	3	2	Q.86
2	L	LL	PT	2	2	Q.86
2	L	LL	PT	3	2	M.75
2	L	LL	PT	1	2	M.69
2	L	LL	PT	1	2	Q.86
3	L	LL	PT	1	1	K.64
4	L	LL	PT	2	2	Q.86
4	L	LL	PT	2	2	Q.86
4	L	LL	PT	2	2	Q.86
4	L	LL	PT	2	2	Q.86
4	L	LL	PT	3	2	Q.86
4	L	LL	PT	2	2	Q.86
4	L	LL	PT	3	2	Q.86
4	L	LL	PT	4	2	Q.86
4	L	LL	PT	1	2	Q.86

Annex E

Subgroup Discovery output tables

Table E.1 - Subgroup discovery output for network number 25, with subgroup identification.

Nodes ID	Subgroup	Target Class	Description
7		2	
5		3	
4		2	
6		2	
11		2	
8	1	1	NACE Div = A.01
9		1	
3		2	
10		1	
2		2	
1		3	
7		2	
5		3	
4		2	
6		2	
11		2	NACE Div = A.01
8	2	1	+
9		1	Type = L
3		2	
10		1	
2		2	
1		3	
7		2	
6		2	Size Class = 2
4	3	2	+
3		2	NACE Div = A.01
2		2	
7		2	Size Class = 2
6		2	+
4	4	2	NACE Div = A.01
3		2	+
2		2	Type = L

Table E.2 - Subgroup discovery output for network number 48, with cluster identification.

Subgroup	Nodes_ID	Target Class	Description
1	1	2	Size Class=2
	2	2	
	3	1	
	4	2	
	5	1	
	6	1	
2	1	2	Size Class=2 + Country Code=PT
	2	2	
	3	1	
	4	2	
	5	1	
	6	1	
3	7	1	Size Class=2 + Country Code=PT
	8	4	
	9	1	
	1	2	
	10	1	
	11	2	
	2	2	
	3	1	
	4	2	
	12	2	
	13	2	
	14	4	
	15	4	
	16	1	
	5	1	
	17	1	
	18	1	
	6	1	
	19	1	
	20	1	

Table E.3 - Subgroup discovery output for network number 51, with cluster identification.

Nodes ID	Subgroup	Target Class	Description
14	1	2	Country Code = PT
13		4	
12		4	
10		4	
1		2	
21		2	
7		5	
16		2	
4		5	
6		2	
9		5	
3		3	
5		5	
15		4	
19		3	
20		1	
17		4	
2		5	
11		3	
8		4	
18		1	
14	2	2	Size Class = 1
12		4	
6		2	
20		1	
11		3	
18		1	
14	3	2	Country Code = PT + Size Class = 1
12		4	
6		2	
20		1	
11		3	
18		1	

Table E.4 - Subgroup discovery output for network number 32, with cluster identification.

Nodes ID	Subgroup	Target Class	Description
5	1	1	Size Class = 1 + Country Code = PT
10		1	
7		2	
19		2	
11		2	
14		1	
18		1	
21		1	
5	2	1	Country Code = PT
17		1	
21		1	
10		1	
4		4	
7		2	
19		2	
11		2	
14		1	
18		1	
13		2	
2		1	
15		2	
3		4	
22		2	
6		2	
9		1	
8		3	
12		3	
20		1	
5	3	1	Size Class = 1
1		99	
10		1	
7		2	
19		2	
11		2	
14		1	
18		1	
16		2	
21		1	
5	1	1	NACE Div = K.64 + Country Code = PT
18		1	

Table E.5 - Subgroup discovery output for network number 21, with cluster identification.

Nodes ID	Subgroup	Target Class	Description
1	1	5	NACE Div = C.10
2		5	
3		5	
4		1	
5		4	
6		5	
7		5	
8		5	
1	2	5	NACE Div = C.10 + LForm = LL
2		5	
3		5	
4		1	
5		4	
6		5	
7		5	
8		5	
7	3	5	Country Code = ES
8		5	
1	4	5	Size Class = 5
9		5	

Table E.6 - Subgroup discovery output for network number 41, with cluster identification.

Nodes ID	Subgroup	Target Class	Description
4	1	1	NACE Div = K.64
3		1	
4	2	1	NACE Div = K.64 + Country Code = PT
3		1	
3	3	1	Size Class = 2
1		2	
2		3	
5		2	
6		2	
7		2	
3	4	1	Size Class = 2 + Country Code = PT
1		2	
2		3	
5		2	
6		2	
7		2	

Table E.7 - Subgroup discovery output for network number 50, with cluster identification.

Subgroup	Nodes_ID	Target Class	Description
	1	99	
	2	2	
	3	3	
	4	4	
	5	4	
	6	2	
	7	2	
	8	2	
	9	3	
	10	3	
	11	2	
	12	2	
	13	3	
	14	2	
1	15	2	NACE Div = Q.86
	16	2	
	17	2	
	18	3	
	19	2	
	20	2	
	21	2	
	22	2	
	23	3	
	24	2	
	25	3	
	26	2	
	27	2	
	28	2	
	29	2	