NOVA IMS Information Management School

MDSAA

Master's degree Program in

Data Science and Advanced Analytics

Providers' Network of a Health Insurance Company: A Social Network Analysis Experiment

Rita Ilda da Silva e Ferreira

Internship Report

presented as a partial requirement for obtaining the Master's Degree Program in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

PROVIDERS' NETWORK OF A HEALTH INSURANCE COMPANY: A SOCIAL NETWORK ANALYSIS EXPERIMENT

by

Rita Ilda da Silva e Ferreira

Internship report presented as a partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a Specialization in Business Analytics

Supervisor: *Prof Doutor* Roberto Henriques

External Supervisor*:* Paula Cristina Dionísio Constantino Santos & Jorge Abreu

February 2023

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, February 27, 2023

DEDICATION

Due to all these people's support, I could never leave this section blank.

As such, I would like to dedicate this work to Cidália Ferreira, my mother, and inspiration. I love you and thank you for all you have given me and all those sacrifices you made to ensure I always had the best.

Likewise, I would like to make a special dedication to Jorge Ferreira, my father, for always educating me by example, teaching me how to be successful in life, and always being present when needed. Also, some of the greatest moments we spent together and those to come will always stay in my memory.

ACKNOWLEDGEMENTS

In light of the importance of certain individuals in concluding this internship report, I would like to list some acknowledgements and give my thanks for all of their support.

To my Nova IMS teachers and the university teaching methodologies, for all the tools that provided me, allowing me to proceed with this project and engage in the professional world, thank you.

To the company that has accepted me as an intern and given me a platform to follow this project, I appreciate it.

Finally, I want to thank my mother for motivating me to write this Report and hence completing my master's degree program and my father for always allowing me to continue with my studies and supporting me in such journey.

ABSTRACT

Despite the large amount of claims data stored by a Portuguese health insurance company, little is known about their in-network of providers and how they relate to each other. The purpose of this work was to leverage the business' management of providers by studying the network of interactions between physicians who prescribed health services (*MCDTs*) cardiology-related to their patients and the health units that executed them, determining who are the most relevant elements, connections, and niches, but also identifying potentially biased relationships and having an overview of the overall topology of the network. Previous research in this field has relied on Social Network Analysis (SNA) to exploit the society mirrored by these bipartite structures – this case was no different, the extensive graph theory provided by this process presented it as the most complete to obtaining the intended knowledge. The methodology followed consisted of a combined approach between insights gathered from three main perspectives: the direct analysis of the bipartite graph, and the study of the dual projection of the network (one on each bipartition) into weighted unipartite networks and unweighted unipartite ones, obtaining more robust results to the limitations of each of the techniques. This investigation included the identification of the most relevant providers through centrality measures, other complementary node and network level conclusions, flagging of several relationships out of the development of two metrics, and several communities detected. This report is part of a growing field in the identification of patterns in a defined population with two types of actors. By applying various techniques and algorithms that considered the networks' unique properties and dynamics of a health insurance/healthcare environment, this report will help to shape future research on related problems.

KEYWORDS

Social Network Analysis; Health Providers; Bipartite Graph; Centrality Measures; Health Insurance; Two-Mode Projection

Sustainable Development Goals (SGD):

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

SNA Social Network Analysis

MCDTs Meios Complementares de Diagnóstico e Terapêutica, meaning Complementary means of diagnosis and therapy

Provider *req.* Provider that ordered the *MCDT*

Provider *exec.* Provider that executed the *MCDT*

VNA Visual Network Analysis

LPA Label Propagation Algorithm

BRIM Bipartite Recursively Induced Modules

NMI Normalized Mutual Information

SAS EG SAS Enterprise Guide

1. INTRODUCTION

In recent years, the term "Social Network" is part of our lives. Most people associate it with the internet and its popular social networking sites. However, it has a wide range of meanings and applications, one of which is Social Network Analysis.

Social Network Analysis (SNA) is the study of relationship patterns among entities, individuals, groups, agencies, or organizations (Hawe et al., 2004). Thus, the network structure and properties, its characteristics, as well as those of its members, have a substantial impact on the outcome of interest.

For several decades and having roots in graph theory, SNA has been studied and applied, showing its importance in mapping the flow of products, services, and information between the network units, as well as processing large amounts of relational data. Tracing back to the 1930s, with Moreno's sociometric approach, some origins of SNA can be revealed. The once-called *sociograms* were represented by nodes and their relations in the form of lines (Oliveira & Gama, 2012). The use of systematic social network analysis was later expanded by other scholars.

SNA-related methodologies are increasingly being used in fields such as psychology, business organization, and electronic communications (Serrat, 2017). Though SNA is also making inroads into the field of health research and health insurance, there is still a significant amount of learning potential, this is one of the motivations for the present work.

It has been shown that medical care service providers can be organized into a social network (SN) that, when analyzed properly, can help explore influential elements, reveal medical patterns of care and behavior, or atypical ones, along with finding relevant groups of actors.

In this document, an SNA internship project undertaken in a Portuguese Health Insurance Company will be presented, along with all the main steps from data collection to results analysis and conclusions, methodologies, tools, and obstacles, among others.

The main objective was to capture the relationships amongst healthcare providers, in particular between the physicians who prescribe exams/healthcare services (complementary means of diagnosis and therapy - *MCDTs*) and the providers that execute them, based on claims data, while studying the feasibility of applying SNA techniques in the company's environment, given all its software/technologies constraints and knowledge limitations in the field. In short, the purpose was to obtain network insights, find influencers, and detect communities inside the network.

Service providers are daily intervenients in the activity of an insurance company. However, during the claiming process that is the only view of the network available. When it comes to the physician that triggers the claiming process by prescribing a service or treatment there is no information (although there is data available). Relationships between physicians and health providers, linked by common patients' claims can be both good and bad, which makes this lack of knowledge a problem. On one hand, there could be biased connections just because the physician works with the service provider. On the other hand, some links among providers can show themselves advantageously for the insurance company, having potential for new partnerships. Hence, this was another of the motivations for the kickoff of this project.

In business terms, the goal was to obtain knowledge, understand and improve the firm's management of providers inside their network, specifically in the branch of medicine of Cardiology, leading back to the analysis of possible clinical protocols, the identification of potential atypical physician-provider relationships and finding patterns, niches, and the most important actors and relationships within the network.

To address this proposal the following research questions were the baseline: I) What are the most important providers in the network? And the most relevant physician-provider connections? II) Is it possible to use Social Network Analysis (SNA) in the company? And in the study of connections between its Providers? III) Are there biased doctors and atypical relationships? IV) Is it possible to detect providers' communities?

To achieve results, the Cross-Industry Standard Process for Data Mining (CRISP-DM) was the process followed. This comprises a cycle of six stages, which starts with Business Understanding, Data Understanding, Data Preparation, and Modeling, ending with Evaluation and Deployment (Azevedo & Santos, 2008).

Although fundamentally important for this work to understand the underlying theory, it is also relevant to address the software/tools that allowed to test and obtain the main findings. These comprise *SAS Enterprise Guide, Jupyter Notebook, Spyder, Neo4j Desktop,* and *Power Bi Desktop*.

Considering the complexity of the network of providers of this health insurance company, new methodologies that came with SNA were needed to examine it and obtain knowledge. As such the main contribution of this work was a set of methodologies, techniques, and even drawbacks that helped uncover the several connections, elements, and patterns of this system.

The following sections of this introduction detail some of the points mentioned above, namely, the context of the internship that supported this research, the specific business and data mining goals, and the project's situation assessment.

1.1.1. Background

Firstly, and as mentioned before, the project was undertaken in Médis, a company that operates in the Portuguese health insurance market since 1996. The company's proposal is to provide its customers with a personalized health service, protecting them through a personalized health insurance plan, and giving them access to an integrated system that includes medical assistance, flexibility, and several programs, among others (*Sobre Nós | Médis*, n.d.). But fundamentally, customers pay premiums in exchange for insurance protection. Médis is a brand of Grupo Ageas Portugal, one of the leaders in the Portuguese insurance sector, which operates since 2005. (*Marcas Do Grupo | Grupo Ageas Portugal*, n.d.).

The department where the project was carried out was Pricing and Advanced analytics. This team comprises both the pricing area (composed of actuaries that study the values of the premiums to charge, monitor the business, and collaborate on product development) and the advanced analytics area (composed of data scientists allocated to business projects, involving several areas of the company, from clinical studies and claims analysis until customer related data examination – through data exploration and data science techniques).

My responsibilities inside the company during my internship of twelve months shifted a little bit. Firstly, I was assigned to a project involving an emerging program: Médis Active. I was responsible for detecting several issues with user data synchronization as well as value's inconsistencies, working closely with the teams involved to resolve them. However, this dictated the end of my project and the beginning of a new one: The Analysis of Health Insurance Claims through SNA – which is the focus of the current manuscript as it will be explained.

1.1.2. Business Objectives and Success Criteria

As presented in 1, the objectives defined together with the company can be systematized as:

- Obtain knowledge, understand, and improve the firm's management of providers inside their network (cardiology related through prescribed *MCDTs*)
- The identification of potential atypical physician-provider relationships
- Find patterns, niches, and the most important actors and relationships within the network
- Capture the relationships amongst healthcare providers, specifically between the physicians who prescribe exams/healthcare services (*MCDTs*) and the providers that execute them

While developing a project of this kind inside a big corporation is important to define what determines a successful outcome. Given that this was the first time that an SNA was being employed inside the company, the time constraints, and tools limitations, jointly with the different stakeholders of the project it was defined business success as obtaining any knowledge that can help improve the firm's management of providers.

1.1.3. Data Mining Goals and Success Criteria

Once the business objectives were defined, it was needed to translate them into data mining goals, which in the later steps played a huge role in the elaboration of the project plan and in determining the analytical direction to follow. The analytical objectives were the following:

- Construct a graph from the data available
- Extract insights on the network structure through statistical measures applied both on node and network levels
- Map relationships between network actors
- Perform community detection
- Explore the viability of the available tools and technology

Through an analytical interpretation of the business success criteria, the data mining success criteria were defined as being able to construct a bipartite network from the available data, and confirm and identify important actors and links in the network. Important to note that this is a project out of the standard data science scoop, so it was harder to define specific numbers as benchmarks for evaluating the success of the project. However, in future projects, this analysis can serve as a baseline to evaluate posterior results, and thus define a success criterion.

1.1.4. Assessing the Situation

At this point, several questions were made to assess the stage at which the project was and help trace the path to follow.

In terms of resources, besides the available data sources (presented in 3.2), and the project support team, the following tools were accessible:

- SAS EG - Used for manipulation, joining, filtering, and transformation of the data
- Jupyter Notebook - Through python, to create the edge list, build the network, calculate metrics, and apply community detection algorithms (main package*: NetworkX* library)
- Spyder - Through python, to develop libraries of functions used in the main notebook script
- Neo4j Desktop (community edition) - Graph data platform explored regarding its potential in the current project and upcoming ones
- Power Bi Desktop – For comparison and analysis of the results of the community detection algorithms, and further conclude about the network of providers

One thing that was important to define from the start, due to the scheduling constraints and this being a new field of investigation within the company, was a contingency plan for the several existing risks. In sum, the two main risks included the tools and software constraints which given the great volume of data, could become an issue, and as mentioned, the time constraints which could mean not finishing the project in time. To safeguard this possibility, two contingencies were defined. First, if it was not possible to perform the data modeling due to either of the risk factors, the resolution would be to focus on only one medical specialty (of the requester provider), mainly, Cardiology (which was what happened, as previously advanced) as for business purposes was the most valuable to analyze (although not being the most profitable one) and it comprised a manageable data volume. Secondly, in case the previous solution was not enough, in agreement with the project sponsors further steps and approaches would be decided.

Also together with the project management team, the work assumptions were clarified, mainly:

- Consider only 2021 claims (with the possibility of, in the future, extending the analysis to other time windows)
- Consider only expenses classified as *MCDTs* (*Meios Complementares de Diagnóstico e Terapêutica*, which means Complementary means of diagnosis and therapy). However, disregard the ones associated with emergencies, surgeries, and hospitalizations – as these situations are out of the standard, and where the concepts of “prescribed” or “referred” applied to doctors and service providers differs from the defined for the project
- Consider only Claims with outpatient coverage (“Ambulatório”)
- Disregard *Covid-19’s* tests related claims – there were a lot, which could create bias in the analysis
- Ignore records related to already discontinued lines of business (“Grupo CTT” and “Subsídio Diário”) – insights that could be found related to these groups are of no value to the company
- Study only claims whom providers belong inside the Médis in-network of providers (the available information on external providers is scarce)

2. LITERATURE REVIEW

In this section, all the key concepts for the development of this project are explained according to the existing literature. It begins with an overview of Social Network Analysis, followed by the study of Bipartite Networks and their shortcomings. Besides, the network assessment using statistical measures and visualization techniques will be clarified. Moreover, an outline of Community detection is provided, understanding how to approach and interpret it. In addition, the Health Insurance Industry is contextualized since this setting is where the research was established. Finally, tools of value to a data science project are approached.

2.1. SOCIAL NETWORK ANALYSIS (SNA)

Throughout this part of the literature review is important to have the concept of graph theory acknowledged. In line with (Bender & Williamson, 2010), graphs (or networks) are one of the main themes of interest in discrete mathematics, being described, in simple terms, as groups of vertices linked by edges. Definitions of graphs may vary between undirected simple graph, directed simple graph, undirected multigraph, directed multigraph, and other notions. Thus, the study of graphs is defined as graph theory.

Likewise, (Serrat, 2017) puts the concept of Social Network Analysis straight with few words. According to this author, this study aims to examine the networks with an in-depth view of the actors and their relationships. In other words, and referring to the views of (Goldenberg & Aviv, 2019), SNA consists of the study of social structures through the exploitation of networks and graph theory.

In a more detailed analysis, Social Networks (SNs) are collections of nodes connected within themselves by relationships (of several alternative forms). These nodes represent actors, such as individuals, teams, social beings, and organizations, whose type varies according to the social context of the network. In graph theory, as previously revealed, nodes are also known as actors or vertices, and the relationships are defined as ties, links, edges, or even arcs (Yang et al., 2017). Contributions of (Oliveira & Gama, 2012) and (Haythornthwaite, 1996) also suggest similar definitions for the previously stated terminologies. Figure 2.1 resembles the notion of a graph (Telatnik, 2020).

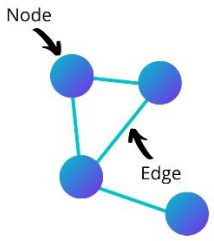

Figure 2.1 – Graph's nodes and edges [Source: (Telatnik, 2020)]

A network as a representation of social relationships within a social context, although with various definitions, dates to the previous century. Earlier contributions of Jacob L. Moreno and Harrison White were keys to the origin of SNA (L. C. Freeman, 2004). As important for the theoretical foundation for this field, was the early work of sociologists like Émile Durkheim and Georg Simmel.

Summarizing, in the first stages and dating back to the 1930s, Jacob Moreno published two books, experimentally grounded through data and analysis (in collaboration with Helen Jennings) that led to the first usage of the expression "network" within nowadays' definition and to the development of the "sociometric approach" (first named "psychological geography") that fundamentally intended to understand groups and the placement of individuals within them (L. C. Freeman, 2004).

During these times, Harvard academics' work also stood out, with collaborations of George Homans, and James Davis, among others. Although, after 1940 and for a period of thirty years there were no major research advances in this area, examples of authors with recent research in groups and positions are Alba, Freeman, Lorrain and White, and Win-ship and Mandell (L. C. Freeman, 2004).

2.1.1. Graph Theory Formulation

Even though the concept of a graph was already defined, it is valuable to formulate it. Both (Bender & Williamson, 2010) and (Oliveira & Gama, 2012) formally define a simple graph **G**, as: $\mathbf{G} = (\mathbf{V}(\mathbf{G}), \mathbf{E}(\mathbf{G}))$**,** where $\mathbf{V}$ is the finite set of vertices of G, and $\boldsymbol{E}$ the set of edges, composed of subsets of pairs of elements of V, or also identified as subset of $\mathbf{P}_{\boldsymbol{k}}(\mathbf{V})$ with k=2, which in turn refers to all the subsets of V with k elements. Accordingly, $|\mathbf{V}(\mathbf{G})| = \boldsymbol{n}$ corresponds to the order of the graph (vertex set cardinality), that is, the number of vertices it has, whereas $|\mathbf{E}(\mathbf{G})| = \boldsymbol{m}$ is the size of the graph (edge set cardinality, overall number of edges), as pointed by (Csikós et al., 2019).

To understand the construction of a graph, the notion of adjacency matrix $A_{n\,x\,n}$ (square matrix) is determinant as it points out the adjacent pairs of vertices and the ones that are not. In short, both (Aguilar, 2021; Biggs, 1974; Csikós et al., 2019) agree that given the vertices $v_1, \dots, v_n$, of the graph G (in this specific definition, simple graph), the elements A_{ij} of A equal one if there is an edge from v_i to v_j and zero if there is not.

Furthermore, (Guichard, 2022) proposes a subgraph $\mathbf{H} = (\mathbf{V}', \mathbf{E}'), \mathbf{H} \subset \boldsymbol{G}$ **,** if $\mathbf{V}' \subset \boldsymbol{V}$ and $\mathbf{E}' \subset \boldsymbol{E}$**.**

Another concept related to G is walk, which is in agreement with the views of (Aguilar, 2021; Bender & Williamson, 2010; Biggs, 1974) a sequence of vertices and edges of the graph (not requiring the edges to be distinct). On the other hand, a path is a walk that does not have repeated edges (trail), connecting the sequence of vertices which are also not repeated. With directed graphs (defined in the next section) comes the notion of a directed path (or dipath), which is, essentially, a path with the requirement that all the edges point in the same direction.

2.1.1.1. Types of networks

To further understand the network in hands during a Social Network Analysis, this section explores the different types of networks. As pointed out before, when examining SNA, there are differences between undirected simple graphs, directed simple graphs, undirected multigraphs, and directed multigraphs, for example.

Starting by defining a simple graph in line with (Rosen, 2019), this is a graph where no pair of vertices has parallel edges in the in-between and without self-loops (edges starting and ending in the same vertex, or equivalently, without the edge $\{x, y\}\ where\ x = y$). A multigraph is a graph that still does not have self-loops but does have two parallel edges or more (Rosen, 2019). (Guichard, 2022) further explains this concept by adding the notion of condensation. According to this author, it is possible to

obtain a simple graph from a multigraph by eliminating the multiple edges, keeping only one of the edges corresponding to those ending points – condensation.

In a directed graph or digraph, according to (Lehman et al., 2015) at least one of the pairs (x, y) and (y, x), can form an edge of the graph, that is, the edges have an orientation. Other studies also refer to this notion as an oriented graph. On opposite sides, there are undirected graphs that have edges connected by unordered pairs of vertices, as described by (Oliveira & Gama, 2012). Both these types of graphs present adjacency matrices, whose only difference lies in the possibility for non-symmetricity of directed graphs, as supported by (Lehman et al., 2015). In a formal notation and according to (Bender & Williamson, 2010), a digraph, namely the pair $\mathbf{G} = (\mathbf{V}(\mathbf{G}), \mathbf{E}(\mathbf{G}))$, is composed of a set of vertices $\mathbf{V}$ and a set of edges $\mathbf{E}$, where $E\ contains\ \{(x, y)|(x, y) \in V^2\ and\ x \neq y\}$. Important to clarify that this is the definition of a simple directed graph. Simplifying, a directed graph can be simple or multi-directed. In the first, multiple edges are not allowed and there is not any self-loop, while the second can display such characteristics.

Following (Lehman et al., 2015) reasoning for these notions, the concept of degree arises (also analyzed in later sections, with more detail). Generally, the vertex's degree is the number of edges incident in it. In a directed graph, the degree is split into outdegree and indegree, that is, the number of edges going out of or into the node (thus, the total vertex's degree is equal to the sum of the indegree with the outdegree).

(Guichard, 2022) introduces a weighted graph as a graph where each edge has a numerical weight associated. Logically, unweighted graphs do not have any weight associated to their edges.

Another type of graph to consider, within the context of this thesis, is the attributed graph where it is possible to specify attributes both for edges and nodes, refining the relationships with additional information (Fionda & Pirrò, 2013). When running a network of health providers several properties can be included to enrich the various objects.

The analysis of vertices and their relationships brings the concept of neighborhood $N(v)$, which according to (Naduvath, 2017) is the set of vertices adjacent to vertex v, meaning that, the degree of v equals the number of neighbors it has.

Table 1 systematizes several types of graphs, according to their edges and loops (Rosen, 2019).

Table 1 - Types of graphs [Source: (Rosen, 2019)]

Type of Graph	Edges orientation	Multi-edges?	Self-loops?
Graph (Simple undirected)	Undirected	No	No
Multigraph (Undirected)	Undirected	Yes	No
Pseudograph (Undirected)	Undirected	Yes	Yes
Directed graph	Directed	No	Yes
Simple directed graph	Directed	No	No
Directed multigraph	Directed	Yes	Yes

Mixed graph	Both	Yes	Yes

Other types of graphs, although not applicable to the present work, are worth mentioning. Specifically, infinite graphs (has a set of vertices infinite or with infinite edges), null graphs (with empty edge set, that is, without any edges), and trivial graphs (with a single vertex – the smallest one) (Deo, 2016).

Both (Naduvath, 2017) and (Guichard, 2022) define a graph as regular if the degree is the same for every vertex, that is, graph G is K-regular if $d(v) = k \; for \; all \; v \in V(G)$.

Additionally, and in line with the previous authors, a graph (a simple undirected one) is complete when each vertex is connected to every other vertex by a unique edge. Moreover, in a connected graph G there is a path from x to y for any distinct vertices $x, y \in V(G)$, being a disconnected graph the opposite case (Aguilar, 2021).

Finally, one of the most important types of graphs for this project is the bipartite graph. (Naduvath, 2017) presents a graph as bipartite if its set of vertices can be divided into two partitions (bipartition (V_1, V_2)), with no two vertices in the same partition being connected. Similarly, but in line with (Aguilar, 2021) graph G is bipartite if two disjoint subsets V_1 and V_2 of $V(G)$ (and non-empty) make up $V(G) = V_1 \cup V_2$. Moreover, each edge has one end-vertex in each of the two partitions. Figure 2.2 represents a visual example of a bipartite graph (Aguilar, 2021).

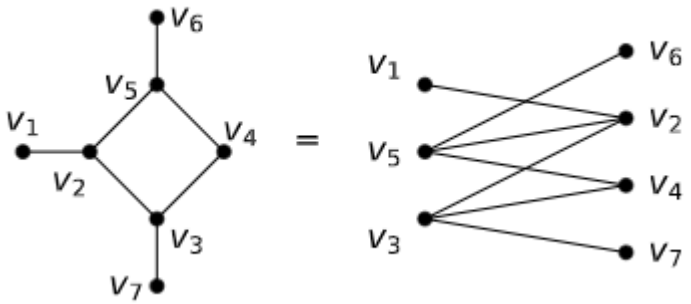

Figure 2.2 – Example of a visual representation of Bipartite graph [Source: (Aguilar, 2021)]

There is general agreement between intellectuals who studied the field under review. So, having defined the basics of graph theory, including some concepts and graph typologies, the following section further debates the topic of bipartite Networks, as in the context of this thesis it is a subject of great importance, with both room for growth and some disagreements in the in-between as to what are the best approaches when following an SNA experiment on elements that form bipartite networks.

2.1.2. Bipartite Networks

One of the reasons why it is important to study the particularities of bipartite networks is the prominence of algorithms, centrality measures, and certain techniques for unipartite networks as opposed to the bipartite case. Thus, it is important to define, firstly, what can or cannot be used and, secondly, what are the alternatives.

This context has been studied by (Aronson et al., 2020; Chessa et al., 2014; Everett & Borgatti, 2013; Latapy et al., 2008; T. Zhou et al., 2007), and both authors agree that there are limitations to these types of structures. (Aronson et al., 2020) argues that many algorithms like *PageRank* or centrality statistics like eigenvector centrality can be deceiving, producing false results – the errors produced when SNA is subjected to bipartite networks are cause for alarm within the scientific community. (Chessa et al., 2014) also approaches these shortcomings but continues adding that, still, it is possible to obtain insights by either projecting the bipartite network into a one-mode network with the awareness of losing relevant information or by following several authors' methodologies and exploiting measures and algorithms applicable directly to the initial network. This first solution is also presented by (Latapy et al., 2008) as an argument for the lack of tools and solutions to approach bipartite networks – two-mode networks do not fulfill the needs for results with relevance, losing clarity and precision. The author's contribution presents methods (as simple statistics) that ease the research in this field, keeping the bipartite form of the network. (T. Zhou et al., 2007) introduces another argument for the lower informativeness of projections – the graph weights in the projection. The author claims that although projecting the network onto an unweighted unipartite network is easier, more information is lost by not considering the frequency of repeated relationships.

As we have seen, several studies have identified the shortcomings of dealing with bipartite networks. However, there has been relatively little literature published on state-of-the-art solutions, or that were implemented - the research to date has tended to focus on the one-mode projections (solution further analyzed in section 2.1.2.2, along with the few examples of novel approaches to this case).

2.1.2.1. Bipartite network basic concepts and properties

A bipartite (or *two-mode*) network $G = (A, B, E)$ is composed of three parts, the two partitions A and B (in this context, also called *bipartitions* (Scheinerman, 2012)) and a set of edges E that represents the links between the two partitions (da Fontoura Costa, 2022). According to the author, bipartite graphs can be directed or undirected although in literature these are commonly presented as undirected – in those cases, each tie corresponds to two directed edges that are going in the opposite direction. Overall, regardless of the graph's orientation, a given edge can go from a vertex belonging to one of the disjoint partitions to the other partition or from the latter to the first, even though in the directed case the origin nodes (starting nodes) and the destination nodes (the end of the edges) are identified.

Thinking on the definition of these networks and referring to the research of (Kunegis, 2015), it is immediate that ordinary networks that contain triangles cannot be bipartite (as the two set partitions need to be disjoint, which does not happen when these three vertices are connected), likewise for undirected graphs with any odd cycle (Bang-Jensen & Gutin, 2009; Lehman et al., 2015), while even cycles are bipartite (Scheinerman, 2012). On the contrary, every tree (connected acyclic graph) is bipartite (Lehman et al., 2015). Accordingly, (Aguilar, 2021; Lehman et al., 2015) stated that being a bipartite graph is equivalent to a) being a 2-colorable graph (after assigning one of two possible colors to the vertices, every pair of linked vertices ends up with different colors), b) there is not any odd cycles within the graph, and c) there is not any odd closed walk.

According to the paper of (da Fontoura Costa, 2022), bipartite networks can be either unweighted or weighted, while the first has binary edges (1 or 0 according to its existence or not), the second has a numeric value associated with the edge.

Revising the algebraic graph theory is important to understand the matrix representation of these networks (*bi-adjacency* matrix) which in turn can be useful, for instance, to count paths (Kunegis, 2015). Appropriately, in a unipartite graph, the adjacency matrix is both square and symmetric, however in the bipartite case the *bi-adjacency* matrix is rectangular - Figure 2.3 illustrates well this type of structure (Ghosh et al., 2010). Taking this object several transformations like matrix decompositions and diagonalizations can be made (Ghosh et al., 2010; Kunegis, 2015).

Vertices	x_1	x_2	x_3	x_4	x_5
y_1	1	1	1	0	0
y_2	1	0	0	1	0
y_3	0	0	0	1	0

Figure 2.3 – Example of a *bi-adjacency* matrix [Source: (Ghosh et al., 2010)]

Some algorithms in the context of two-mode networks include methods to test bipartiteness, algorithms to evaluate if it is possible to transform a certain graph in a bipartite one (through the removal of a set of vertices), and matching algorithms (Eppstein, 2020).

In the field of testing bipartiteness, and as highlighted by (Scheinerman, 2012) graph coloring can be used to determine whether a graph is bipartite, provided that 2-colorable graphs are necessarily bipartite. Following the research of this author, the main idea of graph coloring is to associate to each vertex a color from a palette of k available colors (through a function), so that adjacent vertices have different colors, ending with a k-colored graph. Thus, it is proved that a graph is bipartite if after following this technique (through breadth-first search or depth-first search) in each vertex, using only two colors, adjacent vertices with the same color are never discovered. Regarding the topic of the bipartization of a given graph, it will not be provided a literature review as it is not relevant to the present work, however, more information can be found in the research of (Yannakakis, 1978). In line with the methodologies for testing bipartiteness and performing bipartization, arises the need to develop the topic of measures of *bipartivity*, which was studied by the authors (Estrada & Rodríguez-Velázquez, 2005; Kunegis, 2015). Several examples have been pointed out on its usefulness in determining the degree of *bipartivity*, especially in real-world examples (such as the protein-protein interaction) where knowing how much *bipartivity* exists in a non-bipartite network (cases where the separation into two perfect bipartitions is not feasible) can be determinant (Estrada, 2011). The author also infers that in networks with a community structure, there is also a certain degree of *bipartivity*, which can be handy when approaching a community detection methodology. (Kunegis, 2015) presents four *nonbipartivity* measures, whose algebraic definitions will not be extended further in this paper.

Bipartite networks have shown themselves to be useful in representing certain realities. However, with their analysis also come challenges. (Neal et al., 2022) points its research directly onto the problem (both conceptual and theoretical) that arises from the one-mode projections of these structures. Specifically, ideally the bipartite projections would originate weighted networks, whose analysis is still complex. Another trouble that arises with bipartite networks is their sparse nature, especially in real-world networks (Latapy et al., 2008). (da Fontoura Costa, 2022) agrees with the existence of problems, recognizing the difficulties in identifying repeating patterns under a specific interest, which the author tried to resolve with his work. On the other hand, (Heaney, 2021) argues that SNA experts are already paying attention to the issue, having developed several ways to study it. Nonetheless, with that comes a problem, the variety and flexibility of the methods at disposal bring uncertainty on the approach to follow in each case. (Piccolo et al., 2017) proceeds by mentioning that the mathematical tools are not

as developed to approach bipartite networks as for the unipartite ones, inferring however that direct analysis of the bipartite network can be preferable relative to the projection or even double projection alternatives, although these last can also help. The author calls attention to the importance of redefining measures and methods originally developed for the traditional unipartite structures, to allow a direct analysis of the two-mode graphs. Several reformulated indicators are pointed out as well as who proposed them, examples include centrality measures for bipartite networks, various redefinitions of the clustering coefficient, an emerging clustering algorithm, and other direct analytical methods such as *Galois lattices* and *singular value decomposition (SVD).*

Additionally, several studies have revealed that there are alternative forms of representing two-mode networks. (Kunegis, 2015) refers to the hypergraph, the vector space model, and the projection to a unipartite network (analyzed in detail in 2.1.2.2.).

2.1.2.2. Bipartite network Projection

As acknowledged before, the one-mode projection of bipartite networks is the most widely used approach to obtain insights within SNA. Although some intellectuals stress its limitations, it is still a viable alternative within the context of this project, as such it will be reviewed in this section – defining how to approach it, what are its different types, and what are advantages of this conversion regarding the direct analysis.

A one-mode projection (or equivalently, projection or bipartite projection (Banerjee et al., 2018)) transforms a bipartite network into a unipartite network, that is, the original network structure is projected onto a single node set, over which the analysis pursues (Banerjee et al., 2018; Neal et al., 2022; Piccolo et al., 2017). Given the network's two vertices' partitions, the projection can be made onto the first set where in the new graph only that set of nodes is kept and each pair is linked if, originally, they had common neighbors (from the other partition, naturally), or it can be made onto the second vertices' partition where the rationale is the same (Kunegis, 2015).

(Kunegis, 2015; Newman, 2010) describe the simplest method of projecting the network as being onto an unweighted network, where no weighting is done – the network's topology is not considered nor is the frequency of shared connections. However, (T. Zhou et al., 2007) highlight the importance of having a weighting method when projecting the network – it decreases the gap in structural information relative to the original network. The author follows by indicating that a simple weighting method includes considering the weights of the new links as equal to the number of times the corresponding shared connection occurred. Although (Latapy et al., 2008) partly agrees with the previous statements, they also call attention to a new issue, mainly, the projection onto a weighted unipartite network transforms the challenge of analyzing a bipartite network onto one of analyzing the weighted graph, a topic which is not as developed as the unweighted case.

Figure 2.4 is a good illustration of the projection of the two partitions of the original bipartite graph, whose projection weights are related to the number of common neighbors in the corresponding partition (T. Zhou et al., 2007).

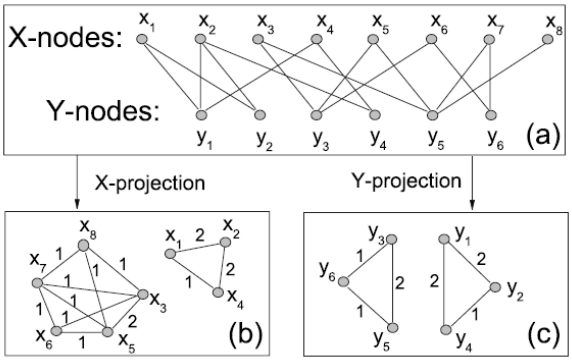

Figure 2.4 – Example of a Bipartite network (a) and its projection onto the partitions X (b) and Y (c) [Source: (T. Zhou et al., 2007)]

(Newman, 2010) states that it is also feasible to project a directed bipartite network, however at the cost of a great complexity increase, so the author does not develop the topic any further.

Reviewing the projections' topic, several methodologies were found in the literature, whose principles vary according to the context and structure of the original network. The contribution of (Banerjee et al., 2018) presents and tests three two-mode projection algorithms whose output is an unweighted unipartite network. The first is the *Brute Force Approach,* which through exhaustion evaluates for all vertices' pairs if they have any common neighbor, building iteratively an adjacency matrix for the new graph. The second is the *Clique Addition Approach* whose basis is the clique concept and its implications on the existence of a common neighborhood. And, finally, the *Matrix Multiplication Method* obtains the adjacency matrix of the new graph through iterative matrix multiplications of the original bi-adjacency matrix. Being the outcome the same (an unweighted unipartite graph), in these cases what varies is the performance of the different algorithms (the proposed output is the same for all the algorithms, just different methodologies). The first case is more time-consuming while for sparse bipartite networks, the two last approaches are better.

The paper of (T. Zhou et al., 2007) speaks to the importance of understanding what is the best weighting method under each circumstance, pointing towards it as a key question within the topic of projections onto a weighted unipartite graph. (Fan et al., 2007) also underlines the importance of weight in the organization of communities, particularly in dense networks. Taking a closer look at possible weighting methods which enable the two-mode projection to derive a weighted unipartite graph from the original one, it can be highlighted in the research of the previous authors and from (Newman, 2010) what is described by some as one of the simpler weighting methods: consider the edge weight as the number of shared connections between the vertices, method that although does not fully captures the original network's information, is an enhancement from the unweighted alternative. Several authors approach alternative weighting methods. The research of (Chessa et al., 2014) obtains the weighted projection by calculating the cosine similarity between pairs of vertices. While (T. Zhou et al., 2007) presents a resource allocation-based weighting method (by taking the degree of resource allocated between objects).

Many authors that revised and applied this topic in literature, such as (Banerjee et al., 2018; Everett & Borgatti, 2013; Piccolo et al., 2017), explained the main cost of projecting a bipartite network - information loss - by only considering one of the node's sets which, in turn, results in a network that is

not an accurate representation of the bipartite network's characteristics. According to (Everett & Borgatti, 2013; Pavlopoulos et al., 2018; Piccolo et al., 2017) one strategy that slightly mitigates the information loss that comes with projection, is the double projection (also known as dual projection) where the results are two networks, each projected onto one node's set. Moving forward, those are explored individually and compared, obtaining a more robust analysis of the original network. The research of (Chessa et al., 2014) also elucidates on the loss of information, including that there is also some controversy among intellectuals as to if a projection can be used to follow a community detection – some authors defend unreliable results, while others as (Everett & Borgatti, 2013) argue that using multiple projections is reliable and can still be used in the proper circumstances. The paper of (Melamed, 2014) promotes a dual-projection method as a way of revealing the community structure of the original bipartite networks.

Although with some backbones, one-mode projections are recurrently used in research and helpful when a direct analysis of the bipartite network is not possible, so it is also important to stress the advantages of following this methodology. According to (Banerjee et al., 2018) an advantage of projecting a given network is that on one hand, it tackles the deficiency of analytical tools that deal with bipartite structures, and on the other hand it enables a faster analysis by reducing the level of complexity of the network.

2.1.3. Statistical Measures

As we have seen throughout this work, the goal of a social network analysis is to produce valuable insights, for example, on the behavior of the featured social system. A statistical analysis of the networks under review is a step towards it, by helping to explore the different roles of the network elements. Some of the key network statistics are regarding the size, cohesiveness, and centrality of the network, which according to (Oliveira & Gama, 2012) give us knowledge without the need for visualization. Besides, a considerable amount of literature on graph theory has been published, exploring the different measures, statistics, and popular metrics and practices. These measures will be seen shortly divided into two parts: actor-level statistical measures (individual-level) and network-level statistical measures, which correspond, respectively, to a description of small units such as the nodes of the network, and the network as a whole.

Following actor-level statistical measures, numerous studies proposed centrality (or prestige) measures as a way of capturing the importance, influence, and role of each node, identifying four prominent centrality measures: degree centrality, eigenvalue centrality, betweenness centrality, and closeness centrality. According to Freeman in 1979, these have been proposed throughout the years as a result of the lack of unanimity on an exact definition of centrality (conceptually and in terms of procedure). (Appel et al., 2018; Goldenberg & Aviv, 2019; Hawe et al., 2004; Newman, 2010; Oliveira & Gama, 2012; Ostovari et al., 2018) suggest centrality as an indicator of where an actor stands (network position) within the social network's structure, reflecting their influence and one sort of "importance". Despite the sociological origins of centrality, the definition of "importance" is the source of some discord among intellectuals varying according to the different centrality measures (Newman, 2010).

(Newman, 2010) describes degree centrality as possibly the simplest centrality notion but still highly instructive, providing information on the connectivity or popularity of a given node (Bloch et al., 2016), the higher its value the greater the vertex's influence over its neighbors (Rochat, 2009). According to

the author and in line with the concept of vertex's degree overviewed in section 2.1.1.1, degree centrality corresponds to the number of connections a node has, in other words, the number of edges incident upon it or, recovering the neighborhood concept (Goldenberg & Aviv, 2019), the number of the node's neighbors. In short, there are several ways of computing the degree, two of which are: through the adjacency matrix and the node's neighborhood (Oliveira & Gama, 2012). Also, according to (Bloch et al., 2016) the degree centrality can be normalized, which in graph G and given the node i with degree $d_i(G)$ corresponds to $c_i^{deg}(G) = \frac{d_i(G)}{n-1}$, where $n-1$ is the maximal possible degree. Initially proposed by Newman applied to unweighted networks, degree centrality (as well as the betweenness and closeness centralities) was in more recent years adapted to weighted networks by Brin and Page (Oliveira & Gama, 2012). In sum, the degree of weighted networks (strength) is calculated as the weights' sum of all the ties that are linked to the vertex. In directed networks, this notion is also known as prestige and it splits into indegree (support is its type of prestige) and outdegree (influence is its type of prestige), which are helpful in the right situations (Newman, 2010; Oliveira & Gama, 2012). Accordingly, the higher the outdegree the more central the node is (synonym for "choices made"), and the higher the indegree is the more prestigious the node is (synonym for "choices received") (Du, 2018).

Degree centrality also enables a group or network-level centrality analysis (degree centralization of the graph), if one summarizes the degree distribution through its mean or variance (Du, 2018).

One of the main limitations of the degree is not considering the overall structure of the network, ignoring the network's architecture and a node's position within it, both of which could provide crucial information (Bloch et al., 2016; Oliveira & Gama, 2012).

Later, in 1987, eigenvector centrality was proposed by *Bonacich,* although there were previous uses (mainly by *Edmund Landau's* paper in 1985 (Holme, 2019)). This concept, alternatively known as *eigencentrality* or *prestige score*, measures the reachability by determining the effect that a node has within the network (Bloch et al., 2016). This notion is an extension of the degree centrality with the advantage of accounting for both quantity (number of neighbors) and quality of the connections – the relationships are not equally important, that is, not all neighbors are equivalent – whereas degree only accounts for quantity (Du, 2018; Newman, 2010; Oliveira & Gama, 2012). Consequently, relying on the notion of prestige, a given node's position is influenced by its neighbor's position in prestige (being, thus, a relative centrality score) (Bloch et al., 2016). In practical terms, the eigenvector centrality can be calculated using the adjacency matrix - be it A of elements a_{ij} the adjacency matrix of G (an undirected graph). Under the assumption that there is proportionality (by the positive factor λ) between a given node's centrality and the sum of the one from its neighbors, the *eigencentrality* is defined by $\lambda c_i = \sum_j a_{ij} c_j$, being c_i the centrality of vertex i and, logically, c_j the centrality for a given node *j* (Bloch et al., 2016; Newman, 2010; Oliveira & Gama, 2012). Subsequently, high centrality values can be interpreted as a node's consequence of having many or/and important neighbors (Pelechrinis, 2015).

Like the degree centrality, eigenvector centrality can also be normalized. Though, when analyzing it, is given preference to the ranking of centralities regarding their absolute values. Nonetheless, this normalization can be done by imposing that the centralities sum to n.

Research has shown that although, theoretically, eigenvector centrality can be calculated for both undirected and directed networks, in the first case the accuracy is greater, while in the latter there are some shortcomings (Newman, 2010). Further explaining, (Newman, 2010) argued that zero-degree nodes have zero eigenvector centrality (and, consequently, so the nodes they are directed to), as such this measure is not frequently used in directed graphs – basically, according to the author's explanation, the standard definition of this concept only works in directed networks when those are cyclic. (Du, 2018) refers the Google's PageRank (a variant of the eigenvector centrality) as an alternative for directed networks, whose adjacency matrix follows the convention $A_{ij} = 1$ given that there is a link from j to i. (Bihari & Pandia, 2015) shows that it is feasible to apply eigenvector centrality in weighted undirected networks by using a weighted adjacency matrix in the calculation.

Proceeding with a completely different way of measuring centrality, we have closeness centrality. This shortest paths-based indicator was initially defined by Bavelas (1950) (Bloch et al., 2016) and measures the distances from one node to the others, which elucidates both the reachability (Oliveira & Gama, 2012) and proximity (Appel et al., 2018) of each actor within the network. Taking the example of a network of people, the lower the distance to other individuals the easier the information is spread (Newman, 2010), thus originally a lower score meant a lower centrality (Bloch et al., 2016).

There are several variations proposed for this indicator. In line with the one from Sabidussi (in 1966), we obtain the closeness centrality by summing the distances from a given vertex to the others, that is, $c_C(x_i) = \frac{1}{\sum_{j \neq i} dist(x_i, x_j)}$ with x_i belonging to the vertices' set and $dist(x_i, x_j)$ being the distance between the nodes x_i and x_j (Bloch et al., 2016; Rochat, 2009). Commonly, the normalized version of this measure, where it is considered the average length of the shortest paths as opposed to their sum, is the one with wider use amongst researchers. Formally, it is defined as $c_C(x_i) = \frac{n-1}{\sum_{j \neq i} dist(x_i, x_j)}$ where n is the vertices' set size (order of the graph) (Rochat, 2009). In this latter version, 1 is the highest possible value of centrality. Under these definitions, only distances are considered regardless of the link's directions from one node to another. However, in the context of directed networks being an incoming or outgoing link matters - this notion cannot be applied to those cases. Regarding weighted networks, some adaptations were developed to ease their use in these situations, such as Dijkstra's algorithm (Newman, 2010).

Another alternative to closeness centrality includes harmonic centrality, which aims to avoid the bias induced by having a few vertices with large distance values (Bloch et al., 2016).

The main drawback of closeness centrality is its restriction to connected networks, and thus the unfeasibility of its application in disconnected networks, as there is not distance set between every pair of nodes (Saxena & Iyengar, 2020). However, although not presented here, within the literature, there are several alternatives to calculate this indicator when a graph is not strongly connected.

Finally, the betweenness centrality will be analyzed. This centrality indicator assesses the importance of a vertex in the task of connecting other vertices within the network (Bloch et al., 2016), by determining how far apart a node is from other nodes in the network (Oliveira & Gama, 2012). In short, it analyzes the extent of control of each actor over the information flow within the network (Ortiz-Arroyo & Daniel, 2010). Relying on shortest paths, for a given vertex in a connected graph, the betweenness centrality corresponds to the quantity of shortest paths that go via the vertex, where the shortest path (graph geodesic) between any two nodes is the minimal number of links between them

(or the sum of the edges' weights for weighted networks) (Newman, 2010). Although with a previous approach by *Anthonisse*, this notion was formally proposed in 1977 by *Freeman* (Bloch et al., 2016; Klein, 2010; Newman, 2010; Zaki & Meira, 2014). Accordingly, it can be defined for vertex v as: $c_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$ where σ_{st} and $\sigma_{st}(v)$ correspond, respectively, to the number of shortest paths from a given node s to node t, and to the number of paths that cross node v (Landherr et al., 2010; Newman, 2010; Oliveira & Gama, 2012). Like other centrality measures, the betweenness can be normalized (in fact, some literature research does not make the distinction, considering the betweenness centrality on its normalized version – (Bloch et al., 2016; Rochat, 2009; Saxena & Iyengar, 2020), for instance). Ranging from 0 to 1, the normalized betweenness centrality can be defined as $\frac{2}{(n-1)(n-2)} \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$ for undirected graphs (Mascolo, 2011).

Values of betweenness centrality often span a large range but compared to the other nodes' absolute values, a vertex with high betweenness centrality can be an indicator of its influence inside the network (as it manages how information is transmitted to the other elements) (Pelechrinis, 2015). Thus, the removal of these vertices can be critical (have the power to disconnect graphs), affecting the communication within the network (the case of a telecommunications network, for instance) (Newman, 2010).

When approaching directed networks, although infrequent, the betweenness centrality can be generalized, by considering the paths' counts in both directions between the vertex's pairs (notice that from node X to node Y and from Y to X the shortest path might be different) (Newman, 2010).

To determine the importance of a given vertex in a weighted network, this definition can be adapted by taking the node strength (adjacent edges' weights sum) (Barrat et al., 2004).

The main limitation of the betweenness centrality is the complexity of its mathematical computation (Landherr et al., 2010).

Following the issue of the applicability of certain measures and algorithms under weighted graphs, (Neal et al., 2022) presents a standard and viable solution: transform the weighted graph onto an unweighted one, through the application of a threshold that determines if certain edges are kept in the new graph (the ones below a certain weight, the weaker links, are removed).

Comparing the previously enunciated centrality measures in line with (Ortiz-Arroyo & Daniel, 2010), while eigenvector, betweenness, and closeness centralities are global measures that account for the overall topology of the network and its information flow, the degree centrality is a local measure. Also, their values do not always point in the same direction, for instance, high eigenvector centrality does not mean necessarily high values of closeness and betweenness centralities (Bihari & Pandia, 2015). In other words, when it comes to centrality rankings across different measures the consistency is not a constant, it can be quite the opposite for some cases – as these measures consider different types of importance and, thus, focus on different topological network properties (Landherr et al., 2010). In his research (Newman, 2010) exemplifies quite well one case where a vertex has high betweenness and low degree (Figure 2.5) - the context of their applications will point to how to approach each case - data from several studies have shown that using different centrality measures frequently results in different conclusions. Also, to apply the different methods, the type of network and study

requirements should be considered as well as the properties one intends to explore (Landherr et al., 2010).

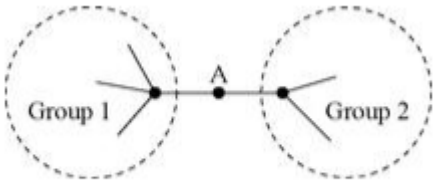

Figure 2.5 – Example of a vertex (A) with high betweenness but low degree centralities [Source: Newman, 2010]

(Ortiz-Arroyo & Daniel, 2010) also agrees, suggesting that to properly evaluate the centralities, the network's flow type for a given application domain must first be identified. Following, Figure 2.6 mirrors this statement by showing the different interpretations of several types of centralities in the context of a graph (Ortiz-Arroyo & Daniel, 2010). Obviously, besides the analysis framework, the robustness of the enunciated centrality measures is highly dependent on the specific topology of the network (Landherr et al., 2010).

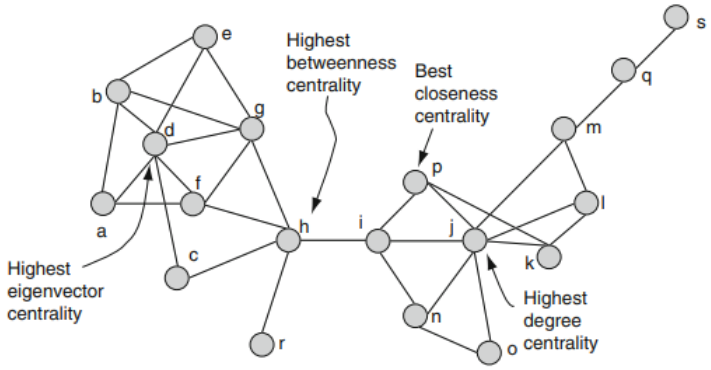

Figure 2.6 – Example of the interpretations of several centrality measures within a network [Source: (Ortiz-Arroyo & Daniel, 2010)]

Furthermore, there are significant variations in terms of the mathematical computational complexity of the different centrality measures. For instance, in degree centrality, the complexity is of $O(n)$ as it is only required to know the number of nodes' neighbors. In contrast, with betweenness centrality (in unweighted graphs) the complexity reaches $O(n \cdot m)$, and $O(n^2)$ with eigenvector centrality, being n and m the number of vertices and edges in the network, respectively (Landherr et al., 2010).

One question that has been raised by some scholars lies in the extent to which a given measure can be compared between nodes in different graphs. The contribution of (Oliveira & Gama, 2012) suggests that it can be done, however when comparing networks with various orders and sizes, one must be aware that some of these actor-level metrics (such as the previously analyzed centrality measures) need to be computed based on their normalized variation.

Noticeable, the previously enunciated centrality measures have some shortcomings (for instance, their unviability for specific types of networks) which lead to the emergence of other centrality indicators

such as harmonic centrality, Katz centrality (a variant of the eigenvector centrality set with *google's PageRank*), information centrality, percolation centrality, flow betweenness, the rush index, cross-clique centrality, and the influence.

As it was shown, determining centralities, and understanding how they work and, when and how to apply their different interpretations, is a topic of great importance for the scientific community and within a social network analysis. Nonetheless, other actor-level statistical measures are insightful, one example is the local clustering coefficient. According to (Newman, 2010) this measure quantifies the level of transitivity in a network. Transitivity, as a property, accounts for the density of triangles in a network (global transitivity) and the density of a node's neighborhood (local transitivity). Thus, the local clustering coefficient, c_i, computed for each node can be defined as stated by (Oliveira & Gama, 2012) as the proportion of vertices' pairs that are neighbors of a particular vertex and are connected by ties, or equivalent as $c_i = \frac{2|e_{jk}|}{k_i(k_i-1)} : v_j, v_k \in N_i, e_{jk} \in E$ (being e_{jk} the tie that connects the nodes v_j and v_k, with k_i as the degree of v_i and $|e_{jk}|$ as the fraction of connections between the nodes within N_i, the neighborhood of v_i) (Oliveira & Gama, 2012).

At a network level, the clustering coefficient can also be calculated in several ways. An elevated clustering coefficient means a network with high transitivity (which in turn tells that the network actors belong to cohesive groups) (Oliveira & Gama, 2012). For the network as a whole, the (global) clustering coefficient as defined by (Newman, 2010) is equal to $C = \frac{number\ of\ closed\ paths\ of\ length\ two}{number\ of\ paths\ of\ length\ two}$, this value converges to one, where a result of C=1 means perfect transitivity, while a value of zero means no closed triads. Another alternative for this calculation is taking the average of all local clustering coefficients for the network vertices (Oliveira & Gama, 2012). Referring to the views of (Newman, 2010), while computing this indicator, it is important to keep the notion that, in a path, direction matters, so it is feasible to apply this concept to directed networks. The paper of (Brinkmeier & Schank, 2005) argues that the clustering coefficient can be misleading without offering further explanation.

To study the global structure of the network, one possibility, as presented before, is the degree centralization of the graph (average degree). This value is maximized in star graphs, where all the nodes are connected to a central node. Also, some adaptions from this notion are emerging, like the Tendency to Make Hub (TMH) which is a recent degree centrality broad measure (global hubness measure), which grows with the emergence of network centrality within the network (Saberi et al., 2021).

In line with (Newman, 2010) and as stated in his research, the nodes' degree distribution is amongst the most important network features. According to the author, to analyze this distribution one can look either at the fraction of nodes or at the total number of nodes that have each level of degree (in the case of extremely skewed distributions the degree analysis on a log-log scale can be useful (Foster et al., 2021)). As different graphs can have equal distributions, this analysis does not capture the full topology of a network. However, it will provide valuable information on the processes that led to the formation of the network and catch a certain amount of the network's structure – usually, in real-world cases, the distribution is more skewed with long tails as most of the nodes have few links (a low degree) and a minority has the majority of the connectivity and reachability of the network (large degree values) (Foster et al., 2021). Both (Foster et al., 2021; Newman, 2010) mention the case of *preferential attachment* where the probability of new nodes that come into the network linking to the

most important actors is high – episode usually associated with scale-free networks, which in turn have degree distributions with high skewness.

Another important statistical measure to analyze the network is density. According to the authors (Newman, 2010; Sahoo et al., 2016), the density of a graph (or connectance) is the proportion of edges present in the graph under the maximum possible number of edges, and it describes the networks' degree of connectivity.

Moreover, the networks' modularity can help measure the network's structure, by assessing how well a network is divided into modules (evaluating how dense the network is within these groups and how sparse is between them), based on the idea that when there is a community structure, the prevalence of edges between groups is less than the randomly expected (Newman & Girvan, 2004; Pavlopoulos et al., 2018). Some community detection algorithms rely on modularity optimization to identify the different groups (Newman, 2010; Zaki & Meira, 2014). This notion will be analyzed in more detail in section 2.1.4 when reviewing community detection algorithms.

Besides, other network-level statistical measures to analyze the whole graph include radius, diameter, eccentricity (Brinkmeier & Schank, 2005), homophily, reciprocity, and average geodesic distance, among others (Oliveira & Gama, 2012).

Taking everything into consideration, and based on the presented measures, one can notice that there is a wide range of options to help explore the network both locally and globally, always considering the different method's constraints, the type of the underlying network, and the application required. Nevertheless, the paper of (Brinkmeier & Schank, 2005) illustrates a network statistic as one that describes the network's key characteristics, can distinguish between various network types, and may be helpful in other applications and algorithms.

So far in this literature review, the arguments were mostly aligned between researchers, although some with a higher agreement rate than others. However, when the topic is bipartite networks the uncertainty and discord are greater. This topic will be reviewed in the following section of this work.

2.1.3.1. Statistical Measures in Bipartite networks

As extensively reviewed in the previous sections, bipartite graphs have their specifications. Although the range of metrics to analyze the traditional unipartite graphs is wider, more diverse, and with a more extensive literature, when it comes to bipartite graphs there are already some viable options (important to notice that in this section the bipartite graphs referred to are unweighted). According to the research of (Pavlopoulos et al., 2018), some analytical techniques for generic networks can also provide valid options for bipartite graphs by making slight adjustments, despite not being specifically designed for them. There is always the option of performing a two-mode projection onto a unipartite network and analyzing the different measures under that structure or even under two projections in the case of dual-projection, that is, two projections, each in one one-mode set (Murphy, 2019) – in fact (Everett, 2016) demonstrates the capabilities of the dual-projection centrality. However, the focus in this section will be both the adapted forms and the bipartite-directed statistical measures.

The degree centrality counts the number of relationships a given node has, to determine its importance. To compute the degree in bipartite graphs, the rationale is equal (Pavlopoulos et al., 2018). However, (Piccolo et al., 2017) advise the normalization of this indicator by the maximum

possible degree for each vertex's set, to tackle the possibility of the two partitions from the bipartite graph having different sizes. In sum, the degree centrality (normalized) can be defined as $\frac{k_i}{|V_1|}$ if i ∈ V_2 or $\frac{k_i}{|V_2|}$ if i ∈ V_1 (being V_1 and V_2 the two disjoint subsets of the graph's vertices, and k_i the node's i degree). From here, the degree distributions can be analyzed the same way as in traditional graphs.

Moreover, the calculation of the closeness centrality according to (Pavlopoulos et al., 2018) is rather complex as this is a path-based algorithm and all paths connecting nodes in the same partition have even lengths. However, (Piccolo et al., 2017) still formally define this metric not mentioning anything related to its constraints. Accordingly, the authors define the closeness as described in 2.1.3, with the difference in the normalized version, where the normalization factor varies according to the partition, mainly, being $c_C(i)$ the node's i centrality, the normalized closeness centrality is equal to $\frac{|V_2|+2(|V_1|-1)}{c_C(i)}$ if i ∈ V_1 or $\frac{|V_1|+2(|V_2|-1)}{c_C(i)}$ if i ∈ V_2.

As previously presented, the eigenvector centrality and thus the importance of a given node depends on the importance of its neighbors. Agreeing with (Piccolo et al., 2017), the definition provided in 2.1.3 can be applied to bipartite graphs being similar to the singular value decomposition of the bipartite adjacency matrix, by imposing that the centrality values cannot be negative. A mathematically rigorous definition will not be provided here, but further details can be found in the paper of (Daugulis, 2012).

Betweenness centrality is a shortest-paths-based method, that when applied to bipartite graphs considers that a node can start and end at each partition of vertices (Pavlopoulos et al., 2018). In detail, (Piccolo et al., 2017) present, like the previously seen measures, the normalized version of this expression because it considers the relative size of the two sets of nodes – thus, the normalization factor proposed corresponds to the largest possible value of betweenness which, in turn, depends on their size.

Although the previous cases can be adapted for two-mode networks, it is important to have some cautions as their initial proposal did not include bipartite networks, which implies that when making the computation each node is handled as belonging to a single mode. Thus, if the intent is to have a differentiation using these measures, a dual projection and separate analysis of each vertex set can be a better approach (Everett, 2016; Everett & Borgatti, 2013; Murphy, 2019).

Besides these popular centrality measures, there are others specially designed for bipartite networks. That is the case of *HellRank*, as proposed by (Taheri et al., 2017). To calculate this measure, it is used the Hellinger distance between pairs of nodes that belong to the same partition. According to it, most central individuals do not necessarily have complete knowledge of the network's structure as only neighbors-based local information is used. However, this new metric has shown a partial correlation with results from other centrality measures' adaptations for this type of network.

To overcome eventual bias when calculating the local clustering coefficient in one-mode projections (from eventually obtaining dense projections from a sparse two-mode network (Latapy et al., 2008)), it is helpful in understanding the density of ties, to calculate this metric on the bipartite graph (or, as an alternative, in the weighted projection, which is also complex). Several redefinitions of clustering coefficient for two-mode networks were presented, including the work of (Robins & Alexander, 2004) where tendencies for bipartite cycles (locally) are measured. (Liebig & Rao, 2014) presents various

other implementations from different authors, referring that most formulated the metric as the concentration of cycles of length 4, as the case of (Zhang et al., 2008). (Latapy et al., 2008) introduces two reformulations of the classical definition of the clustering coefficient, the first (Latapy clustering coefficient) does not include the pairs with empty overlaps, while the second (Robins and Alexander clustering coefficient) earlier proposed by (Robins & Alexander, 2004) investigates the probability of four nodes being connected in a chain, forming a square.

Naturally and like one-mode graphs, by averaging the values of the local clustering coefficient for all the network's nodes, it is obtained the global clustering coefficient which is an overall representation of the clustering across the network.

Other than the global clustering coefficient, there are numerous other network-level statistical measures applicable to this type of graph, (Latapy et al., 2008) point some, including the analysis of degree correlations and degree distributions. (Guillaume & Latapy, 2004) adds the option of also studying the clustering coefficient distribution throughout the network. Furthermore, (Borgatti & Everett, 1997) refer that the degree centralization can be applied directly to the bipartite structures, however, to avoid interpretation issues, the denominator of the formula is redefined using a new maximum. The author also complemented this overview by presenting, in addition to two-mode degree centralization, the two-mode closeness centralization and the two-mode betweenness centralization.

Lastly, there are still some literature gaps on how to approach weighted bipartite networks, especially on the topic of centrality measures and other complementary graph statistics, consequently, this work will not provide further elaboration on the subject to avoid inconsistency with arguments coming from different sources.

2.1.4. Community Detection

Several intellectuals have highlighted the key role of community detection in analyzing the network's structure and acquiring knowledge from them (Newman, 2006; Peng et al., 2021). Community detection is a field of SNA with the objective of perceiving community structures in networks as a way of understanding the underlying society's dynamics and how the flow of information occurs (George et al., 2020).

In recent years, a large variety of community detection algorithms have been proposed. (George et al., 2020) mentions several different methodologies including statistical-based ones, optimization approaches, spectral methods, likelihood-based, and greedy methods (classical), among others.

Starting with the Greedy Modularity Maximization algorithm, it is important to define the concept of modularity in the network's context. (Newman, 2006) formally defines modularity as being the number of edges belonging to a given group less the expected number of edges that would randomly be assigned to that group, meaning that high/positive values suggest the presence of a community structure. The modularity optimization through the Greedy algorithm as proposed by (Clauset et al., 2004) is done at each step, that is, through an iterative process that starts with communities of one element each. At each iteration, communities are merged in a way that optimizes the networks' overall modularity (Newman, 2006). Furthermore, and still about modularity optimization, (Blondel et al., 2008) proposed the Louvain method. This algorithm, like in the previously described case, involves, in

the first phase, the modularity's local optimization at each step. The algorithm's second phase consists in constructing a new network of linked communities and redoing the first step.

Another standard community detection method proposed by (Newman & Girvan, 2004) is the Girvan-Newman algorithm, which is betweenness-based, that is, to divide the network into groups, the edges are iteratively removed based on the highest betweenness measures that are recalculated at each cycle. Ultimately, they also proposed a strength indicator of the number of communities to be obtained. Nonetheless, according to (Vispute, 2020), this method is not the most efficient in large-scale networks.

Next, the Kernighan-Lin Graph Partitioning algorithm (Kernighan & Lin, 1969) aims to divide the network into two equal-sized optimal partitions. The process proposes to make cuts on the edges (which have an associated cost) minimizing the overall cost of doing it.

Additionally, the clique percolation method is also an alternative community detection algorithm. Accordingly, the communities are based on k-cliques, specifically, focusing on locating all the maximal cliques. In the end, the communities are combinations of all cliques of size k that are accessible via adjacent k-cliques (Palla et al., 2007).

Proposed by (Raghavan et al., 2007) the label propagation algorithm (LPA) is an iterative process that at each step, given the neighbors' labels, updates the nodes' labels (synchronously if these are updated in parallel, or asynchronously) until strong communities (groups of nodes with the same labels) are obtained.

When the type of graph is bipartite, exists an agreement between authors as (Larremore et al., 2014; Melamed, 2014), specifically, there are some limitations, and as such, the subject should be undertaken carefully.

The main solutions used to approach this issue include the use of adaptations of unipartite graphs' community detection algorithms or community detection made onto one of the two-mode projections of the original graph (Arthur, 2019). Another alternative shown by (Melamed, 2014) is the dual-projection approach which combines (through the maximization of within-community ties) the results of applying unipartite community detection algorithms to each mode, obtaining an overview of the overall structure of the network (which essentially consists of the combined strategy followed by some researchers, with the advantage of having a criterium to merge the results).

However, when using the projection's alternative, several authors call attention to the possibility of information loss or the non-interpretability of the results, being the use of multiple projections a possible solution (Kernighan & Lin, 1969; Melamed, 2014).

(Liu & Murata, 2009) proposed an adaptation of LPA, based on the asynchronous model, viable for two-mode graphs. Also, an extension of the Louvain method developed by (Peng et al., 2021) can be highlighted. This agglomerative clustering method has its main difference in the modularity expression. Other alternatives, for direct community detection on the bipartite network, include the biclique (extension of the k-clique algorithm) and the BRIM algorithm (bipartite recursively induced modules) (Sun et al., 2017).

Finally, to measure the quality of the community detection results, or to compare different methods, it is useful to have evaluation metrics. However, it is important to distinguish between external and internal criteria. In the first, there are reference communities over which the evaluation can be made. In the second, there is no reference, being a blind quality assessment (Manning et al., 2009).

According to (Labatut, 2015) the most common external metrics for evaluation are purity, Rand index, and normalized mutual information (NMI). Nevertheless, this author also presents the modified NMI.

Concerning internal scores, similar uses in the clustering evaluation can be applied to community detection. In the experimental results of (Vispute, 2020), the performance of different algorithms is evaluated and compared based on the number of communities, modularity, and execution time. Regarding modularity, the author adds that the average of this measure indicates the strength of the communities, the higher the better. Also, the execution time is relevant, especially with bulky data. Finally, the number of communities, when analyzed in parallel with the other two criteria can be significant.

2.1.5. Graph Visualization

In the field of SNA, the topic of graph visualization is important, and when performed effectively it can leverage the analysis. In agreement, (Bastian et al., 2009) refer that this procedure is challenging by nature and requires an adequate methodology. Although the authors focused on the *Gephi* software, some useful techniques were presented like setting up the visualization according to the graph attributes and filtering out nodes and edges to reduce the network and thus concentrate on the most significant features. In the same line of thought, (L. Freeman, 2000) adds that the centrality of this subject in SNA is present since the first contributions, playing a significant part in acquiring insights into the network's structure. On the other side, (Venturini et al., 2021) partly agrees adding that visualization charts are admired, yet they are often feared due to their ambiguity.

Referring to the views of (Venturini et al., 2021) and focusing on the practical component of Visual Network Analysis (VNA), typically the procedure that starts with points and lines includes a connectivity-based nodes positioning, node size defined in relation to their importance, and category-based coloring. Another approach includes force-directed graph drawing which, in simple terms, assigns forces between the set of edges and nodes to define their positioning. This has the advantage of enabling an analogous reading to geographical maps (having a suggestive side), contracting with the hazy and overlapping features of the points-and-lines charts. These authors exemplified VNA by filtering nodes based on a set of attributes, keeping only highly connected actors and their links. Regarding the network's design, they referred to several possibilities like making the nodes' size proportional to their degree, hue the visualization based on a given variable or variables, taking advantage of heatmaps to highlight higher density parts of the graph, and assessing several different layouts.

(Hoffman, 2021) approaches the importance of the layout in network visualization admitting his preference for algorithms such as the *Kamada Kawai* or the *Fruchterman-Reingold* to determine the position of the several nodes. (Heer & Boyd, 2005) presents the option of using a spring-embedding algorithm-based layout (force-directed) which can provide better visualizations by grouping elements based on increased connectedness. Consulting the documentation of the *gephi* software (*Tutorial Layouts Gephi*, 2011) several layouts are presented depending on the feature to highlight, mainly

OpenOrd (focused on the network's divisions), the ones focused on complementarities (*ForceAtlas, Yifan Hu,* and *Frushterman-Reingold*), layouts ranking-aimed (*Circular* and *Radial Axis*), and finally *GeoLayout* that emphasizes the geographic characteristics of the network.

VNA is growing day by day, and with it, several new techniques are being proposed, which highlights the relevance of implementing it in a social network experiment.

2.2. HEALTH INSURANCE FIELD/INDUSTRY

Simply put, health insurance comprehends a contract between two entities: a company and the consumer. In exchange for a premium payment, the company agrees to participate in the payment of the insured person's healthcare costs. (KAGAN, 2022)

One key point to distinguish within this project is the concept of the provider. (*What Is a Provider? | Healthinsurance.Org*, n.d.) defines it as health professionals, or physicians in specific, who provide health care services. Some examples include cardiologists, hospitals, physical therapists, and others offering specialized medical services. A certain amount of insurance health plans have in-network providers and/or allow for out-of-network providers. (*Understanding Health Insurance*, n.d.) refers to physicians and healthcare institutions that provide patient services covered by the insurance plan as in-network providers. Conversely, it defines out-of-network providers as the ones not covered under the client's insurance plan.

In addition, a health insurance claim occurs once the policyholder or provider produces a request for a fee or reimbursement of the costs associated with a given treatment (*Medical Insurance Claim Procedure & How Does It Work?*, n.d.). Accordingly, in a healthcare context and data-driven approach, claims data are digital records of transactions that contain information on patients' and providers' interactions (Burton & Jesilow, 2011). Specifically, this type of transactional data commonly contains details such as the identification of healthcare professionals involved, the identification of the patient, and procedures discrimination (type, timestamp, costs for the patient and company, existence or not of pre-authorizations), among others. (Appel et al., 2018)

Claims data analytics is extremely valuable for the business of health insurance companies. However, it is important to understand in what sense: What is currently being done? What were some of the previous approaches and studies on the subject? (Konrad et al., 2020) states that for research, creating policies, and making decisions, more organizations are turning to claims data. This analytical dependence is in great part due to the analytical developments verified over the past few decades. Considering the interpretations of (Konrad et al., 2020), towards the end of the 1980s, researchers developed models for categorizing individuals based on their levels of risk and thus predicting future costs in healthcare. Following, and throughout the 1990s, academics used this data to study the provider: to understand their services and use of resources. Subsequently, over the last years, there was research on health systems, medical issues, healthcare utilization, forecasting on several matters, and determining fraudulent providers (context studied by (Chandola et al., 2013)), among others. Despite great advances in this field, research on the use of claims data presents some key challenges like the alleged purpose of the data being administrative use and its private property, for instance.

The findings of (Appel et al., 2018) also address the trends in this area. Namely, the analysis of interactions between medical experts. Examples of authors such as (Landon et al., 2013; Mandl et al.,

2014; Sauter et al., 2014) are given, which respectively and from claims data, constructed networks of health providers with common patients, evaluated frequency and unity of certain provider formations, and performed physicians' community-detection.

This concept will be further analyzed over the following segments of this literature review, specifically in 2.2.1. SNA in Healthcare / health-insurance environment, using claims data.

Medical prescriptions are one of the ideas underlying the claims process. These are medical written orders for the patient to receive certain treatments, undertake an exam, or buy a particular medication. Another term connected to this thesis is referral, which is a request for the patient to see a specialist or receive specific medical services from a physician, as stated by (*Referral - Glossary | HealthCare.Gov*, n.d.). The notion of referral, as most of us are aware, is a common practice within the healthcare industry. It can be divided into two types: when individuals, on their own, seek a medical diagnosis or treatment for a symptom from a provider of healthcare, or when it is the doctor who refers the patient to a specialist or another healthcare professional.

Prescriptions are connected to referrals. However, it is important to distinguish them. Taking the example of physical therapy, while the prescription is a written order from the referring physician for the patient to receive treatment, the referral is an authorization/recommendation to visit an in-network specialist (*What Is the Difference between a Prescription and a Referral for Physical Therapy? - ProAction Physical Therapy*, 2021).

2.2.1. SNA in Healthcare / health-insurance environment, using claims data

This part of the literature review will focus on understanding previous research and its limitations concerning SNA in a healthcare or health-insurance environment, using claims data. As shown in 2.2, this subject is becoming an increasingly relevant and extremely important topic for both companies and researchers, being SNA one of the best-emerging tools to approach it.

Further examining the research on this topic, (Mandl et al., 2014) use claims data to pinpoint the network of healthcare professionals involved in each patient's treatment, highlighting their cooperations and partnerships. Differently, but also employing an SNA approach, was the experiment of (Landon et al., 2013). The main purpose was to identify, systematically, physicians who display tight collaborations, flag potential ACOs (groups of providers that voluntarily give care to a specified group of patients), and maintain the care within the network. Comparably, referring to the research of (Appel et al., 2018), the focus was to map data from claims as graphs, focusing on connections between physicians who treated the same patient at some point (during their consultations). The referral concept is adjacent to this last study - although two doctors connected does not imply a personal relationship, it suggests that there is some type of acquaintance or recommendation of patients from one to another, even if it's just because of their similarity of profiles, locations, or specialties.

SNA under the context of healthcare or health insurance is often discussed yet rarely disclosed the why of using this approach as a leveraging tool of the analysis. (Appel et al., 2018) debate this topic arguing that it is an advantage as it presents various methods to measure the relative importance of the nodes. Consequently, instead of trying to analyze, unilaterally, each physician and each of their links inside the network, using traditional transactional techniques as an attempt of finding the ones that are in a central place, this type of analysis eases the process of individualization. The flagging done is either

based on the nodes' relationships, their associations with said "influential physicians", or just by being a bridge that supports the main topology of the network. The contribution of (Scott et al., 2005) also defends that SNA is a tool that when it comes to detecting quantitative characteristics within patterns of interactions in health organizations is a beneficial method, being an upgrade and a good complement of qualitative description. These authors add that SNA enables quantitative comparisons between diverse network architectures and interactions among people using the notion of the sociogram (an illustration of the interactions within a group), stating that SNA has a diverse range of uses, including the support to organizational transformation. (Ostovari et al., 2018) defended in their study (about the identification of key players in the care process of patients) that, given the extension of health data, SNA is a growing method that makes it easier to evaluate the collaboration of healthcare professionals. Adding that this type of research aims on giving healthcare professionals and policymakers tools to help them make well-informed decisions when creating insurance policies and coverages.

When it came to the networks' construction, in most cases some data transformations like filtering were applied beforehand. Whereas (Mandl et al., 2014) disregarded claims involving direct care and the ambulatory care category, focused on the providers with a paper of relevance in the network (more than fifty patients), and grouped some care specialties, (Landon et al., 2013) ignored patients lacking claims data, and merged information on doctors' specialties and regions, from their records of meetings with physicians from a Medicare program of beneficiaries. Also, only the 20% strongest relations between nodes were kept, ending with a sparser network of the most influential relations.

Similarly, both (Landon et al., 2013; Mandl et al., 2014; Ostovari et al., 2018) constructed a network based on the shared patients among physicians. However, in the first case, it resulted from the projection of a bipartite network (composed of patients, physicians, and their relations) into a unipartite one. Additionally, (Ostovari et al., 2018) made the edge weight equivalent to the number of shared patients between providers, and (Mandl et al., 2014) took advantage of the regional features of their universe and, for each region, provider networks were created.

The analytical methodology of (Mandl et al., 2014) was conducted through summary measures, descriptive analysis, and t-tests, as a tool to discuss provider collaboration regarding their cohesion, constellations, and shared patients. (Landon et al., 2013) revealed the community structure of the different networks employing modularity maximation-based algorithms of community detection: Newman's algorithm and the refined version by (Newman & Girvan, 2004) (adjusting the method with a null model to account for the reality that doctors with a high degree are more likely to be connected than the ones with a low degree), and to ease the results' analysis, smaller communities were disregarded, and other filtrations were done. To study the nodes' centrality, (Appel et al., 2018) applied to the network the centrality measures of degree, eigenvalue, betweenness, and closeness, exploring their change over time. To complement, the network density was measured both in the network as a whole and in certain subsets to quantify the connectivity among physicians. This methodology, to understand the importance of health providers, was also part of the study design of (Ostovari et al., 2018), which complemented it by determining the clustering coefficient. (Landon et al., 2013) also calculated several network measures, including the degree (in this context, the number of doctors each one is connected to, being the link shared patients) and the "adjusted degree" whose goal was to take into consideration the volume of patients – the proportion of each doctor's degree under the total of patients associated to the doctor itself.

Different research limitations were presented by these intellectuals. (Mandl et al., 2014) highlighted in their investigation the following: the use of a single data source, leading to missed information on providers' connections through patients not covered by Medicare and, thus, not included (as shared patients, collaborations, etc.) – the selection of the highest rated regions in market share was a way to lessen its effects. Also, by using a subset of the total of patients some conclusions might have been under or overestimated. Similarly, (Landon et al., 2013) affirm that one of the shortcomings of their research is the incompleteness of data – in this case, the records were from a single year and from only one hospital referral region out of fifty-one possible. It is also mentioned that the age of the data (study from 2013 with data from 2006) could be seen by some intellectuals as a problem, although only future work could confirm it. Finally, this paper adds that these techniques may be useful for locating medical groups that have the best chance of offering quality treatments.

After an extensive review of the existing literature on the analysis of providers' relationships within the healthcare system and health insurance industry, with a focused view on the medical prescriptions made by the physicians, it was possible to identify a literature gap/scarcity – it does not mean, necessarily, that it was never implemented or without value for the business (because it has). However, being the health-insurance sector private, existing implementations are not as easily shared.

In sum, most of the studies investigate the subsequent patients' visits from one specialist to another, ending up with a network of connected physicians by common patients – from which is possible to determine groups of communities, physicians' centralities (outlining the relative significance of doctors in the doctors-doctors network), among others. In the case of the current thesis, the goal shifts a little, although it has a lot of common ground. Instead of focusing on consultations and referrals between doctors, the intent is to look at relationships between physicians and health service providers by taking the medical prescription that the physician ordered – one can observe the example of a cardiologist that prescribes the patient a special exam that in hand is performed in a certain health-service provider organization. In short, it is obtained a network of physicians and health service provider organizations, linked by common patients/prescriptions. Also, another gap that was previously noticed, but is important to emphasize in this context, is the literature limitation on how to approach bipartite networks, by having nodes of different types – which occurs in the current project.

Nonetheless, this literature review was a great support for the current research. Firstly, it highlighted the advantages of implementing SNA to gather the knowledge required. Next, it elucidated on possible data preparation steps before the network construction (e.g., filtrations, groupings, assumptions, relevant data to consider), on how to retrieve value and interpret the results obtained with a focus on the field under review (e.g., the notion of referral), it presented several techniques to highlight important actors of the underlying society (e.g., centrality measures) and describe them (e.g., the clustering coefficient), it pointed constraints that this study could face and how to tackle them (e.g., the bipartite network problem, the two-mode projection, and the interpretation of the new network in the healthcare field), it described useful practices (e.g., edges weighting methods), and it showed the research gaps that, on one hand, were a window of opportunity for new findings, and on the other hand, called the attention to certain cautions to have and shortcomings to confront.

2.3. TOOLS AND TECHNOLOGY

In this segment, several tools that can add value to a data science project will be approached.

Starting with SAS Enterprise Guide (SAS EG), this interface to the main application of SAS has the potential of creating a workflow that allows the user to perform analytical tasks and their reporting. Some useful capabilities go from its graphical user interface and its flexibility of accessing the several menus, until its process flows, projects' organization, data querying options, statistical methods, models, data transformation abilities, reporting, security, and performance (Meyers et al., 2009; SAS, 2017).

Python is a dynamic programming language, with an extensive library of resources useful for several types of programming: procedural, object-oriented, and functional (Linner et al., 2021). There are several python compilers like *Jupyter Notebook, PyCharm,* and *Visual Studio Code* (*Top 13 Best Python Compiler For Python Developers [2023 Rankings]*, 2023).

Neo4j Desktop is a graph data platform that includes several components such as analytical tools, graph algorithms, data integration alternatives, data discovery tools, and visualization capabilities (Neo4j, 2018).

Microsoft Power BI is a business intelligence software that can help the process of storing, transforming, analyzing, visualizing, and sharing data (Hart et al., 2023).

The remainder of the thesis is organized as follows. Firstly, the CRISP-DM methodology and its role in the project is explained, along with the project plan's presentation. Secondly, after a deep analysis and understanding of the data in hand, the main data transformation steps are presented as well as their basis regarding the business assumptions and analytical-driven grounds. Following by an explanation, exposition, and analysis of the generated networks and their properties. Finally, the main findings are examined in light of the stated research goals.

3. METHODOLOGY

This section illustrates the process taken to tackle the current project. The methodology followed to help guide it was Cross-Industry Standard Process for Data Mining (CRISP-DM). This was the approach since it is effective, easy to understand, and is not rigid in the sequence of the different stages, in fact, it requires a dynamic transition and alternation between them, as they depend on each other (Chapman et al., 2000). CRISP-DM was developed in 1996 to provide a framework to address data mining projects and comprises six phases (independent of both the industrial context and technology used) (Costa & Aparicio, 2020):

1. Business understanding
2. Data understanding
3. Data preparation
4. Modeling
5. Evaluation
6. Deployment

Within these phases, the model also encompasses several levels of tasks, from wider to smaller ones, mainly generic tasks, specialized tasks, and process instances (Wirth & Hipp, 2000). Starting with Business understanding, the focus is to understand the objectives and requirements of the undertaking both in a business and technical context, outlining an initial project plan (Chapman et al., 2000). Accordingly, and referring to the views of the same authors, data understanding aims on getting to know the data, and its quality, and gather some useful insights. The data preparation includes all the steps that transformed the raw data into the final dataset, this phase is dependent on the modeling, so a back-and-forth might be needed. In the modeling stage, several techniques and models are applied to approach the defined problem. The goal of evaluation is to overview the model and the steps that lead to its creation, understanding if it meets the defined needs. Finally, in the deployment, the knowledge gathered throughout this cycle is put into practice, which can happen in various forms (e.g., implemented directly in the business, or through the elaboration of a simple report) (Chapman et al., 2000). The life cycle of a Data Mining project according to CRISP-DM can be depicted in Appendix A (Dinh et al., 2020).

Also, the diagram presented in Appendix B, provides a better understanding of the several steps followed to embrace this research, which will be complemented with the written explanation over the following sections.

3.1. BUSINESS UNDERSTANDING

Before starting the project, it was important to define and understand the project context, objectives, requirements, and constraints, both from business and data mining perspectives (see subsections of Introduction), making sure that the problem and designed plan were well defined.

3.1.1. Project Plan

Once there was an understanding of the project objectives both from the business and data mining perspectives it was important to outline an initial plan, which naturally suffered some modifications throughout the project's cycle.

Especially for the data preparation and modeling but also applicable to all the other stages, it was important to think of the techniques and approaches that would be a better fit to address the problem defined. After an extensive review, the type of analysis chosen to be undertaken was SNA. The intended was a process that would capture the relationships between providers (the ones that prescribe and the ones that execute), understand how these behaved in the context of prescribed *MCDTs*, but also to identify patterns, important elements, or atypical ones, generating insights for the business that could help manage their network of providers and thus add value to the company. SNA met these requirements, it is a tool that allows studying the structure of complex networks while capturing the social interactions between its elements and quantifying given aspects such as the importance, relevance, and density, or identifying patterns, groups, or specific elements, based on an extensive graph theory and mathematical methods applicable to the most diverse type of networks. Also, it has been demonstrated in previous works the effectiveness of this approach in similar contexts.

Having this point set, the project was defined (Table 2 – Project Plan) using as workflow reference the CRISP-DM methodology, for its easy implementation and dynamics that allow alternating between the several stages. For the business understanding part, the objectives, assumptions, and requirements were defined along with this planning. Further, it will comprise several tasks to get acquainted with the data, extract initial insights, and find eventual quality issues. Data transformation steps were partly defined according to the SNA methodology but will also be based on insights gathered from the previous step, in sum, it will comprise all the tasks that allow transforming the raw data into the final dataset that will be used in the modelling, while also collecting additional knowledge throughout the process. In the modelling, the network of providers will be constructed, and through several SNA techniques, statistical analysis, network visualizations, and community detection, several conclusions and knowledge will be obtained with the goal of meeting the defined objectives. In the evaluation, the results are going to be discussed to determine if the initial criteria for success were met, that is, if the initial proposal was addressed, or if any stage of the project needs to be revisited. Finally, in the business environment, the results will be presented in the format of a report that will be later discussed to determine the next steps to follow.

Table 2 – Project Plan

	1. Business Understanding	2. Data Understanding	3. Data Transformation
Steps	• Define objectives and success criteria (Business and Data Mining) • Acknowledge the resources available • Identify risks • Clarify assumptions • Revise the existing literature on the subject • Determine the methodology to follow • Outline the project plan	• Collect the data • Describe the data (fields, content, among others) • Get the first insights • Data Exploration	• Filter data • Merge different data sources • Query data • Select data • Handle missing values • Perform data aggregations • Understand if any of the risks is verified and act accordingly • Identify additional steps needed • More in-depth data exploration • Final dataset extraction

Tools	--	**SAS EG**	**SAS EG**
	4. Modeling	**5. Evaluation**	**6. Deployment**
Steps	• Report some characteristics of the final dataset • Graph Construction • First network insights • Network Statistical measures (determine providers' and connections' importance, among others) • Address the issue of bipartite networks • Develop additional measures to characterize the providers and their connections • Community detection using several algorithms • Visual Network Analysis (VNA)	• Evaluate the insights gathered	• Deliverables: report and presentation • Define the next steps along with providers management
Tools	**Jupyter Notebook, Spyder, Neo4j, and PowerBI**	**_ _ _**	**Microsoft Word and PowerPoint**

Additionally, throughout this project, several different nomenclatures will be used to refer to the same object, specifically, for the two types of nodes of the bipartite network. Table 3 details them to ensure a better understanding of the current thesis.

Table 3 – Different nomenclatures for each type of provider

Provider Type	Nomenclatures
Requesting Provider	Physician, Doctor, Provider (requester), Provider *req.*, Requester, Requesting doctor
Executing Provider	Performing provider, Provider (executer), Provider *exec.*, Healthcare unit, Service Provider, *MCDTs'* Provider, Executer

3.2. DATA UNDERSTANDING

The first step in this phase of the project was data collection. Afterward, it came data exploration to familiarize oneself with the data – understand the volume of data, find eventual quality issues, get preliminary insights, and plan the upcoming steps.

Several data exploration techniques were followed, mainly, data visualization, analysis of statistical measures (to understand the structure and nature of data), examination of the variables and their connections, and analysis of the distributions of values, which enabled a better understanding of the underlying data. These steps were fundamental for the project. This comes in agreement with previous findings from the scientific community, for example (Sheshnath Jaiswal, 2022) states that the quality of the analysis will increase with the strength of the data's knowledge base.

The business goal was set. However, throughout the project, the work was done closely with the people responsible for the decisions process – as discoveries were made, all parties involved reached a consensus on the best approach to follow. As a result of this dynamic, it was possible to understand what the best sources of data were, right at the initial stage. Being Médis a company that values

information and it being essential for its core business, there were a lot of available sources of information (claims data, customers' personal and consumption data, digital information, clinical data, among others) all stored under the format of relational tables in the company's database (SAS EG).

Naturally, after identifying the raw source tables, a lot of data transformation was needed, to respect the requirements of the analysis, resolve the existing issues, and allow the implementation of the analytical techniques, among others. So, in a way, the data understanding and data preparation phases of this project are connected, complementing each other and having been done iteratively. Also, one important step for this project was to understand what were the keys that would allow connecting the different tables. Next, the different datasets that were used as well as the type of information and fields they contained are presented.

Dataset *DMNL_SINISTROS* (i):

This relational table was fundamental for this project – the main source of information, so a more detailed analysis will be presented. This extensive dataset provides claims data for all Médis customers since the year 2016. Naturally, this data was generated by billing, comprising a wide range of health treatments and services, such as *MCDTs*, outpatient, inpatient, and pharmacy expenses, among others. The main attributes it includes are *Claim ID, Claim Sequence Number, Client ID, Claim Status, Procedure Start Date, Procedure End Date, Claim Type, Policy Start Date, Policy End Date, Provider ID (executer), Provider ID (requester), Total Value, Value, Copayment Amount, Covered Amount, Geographic Location* (of the executing provider)*, Specialty* (of the executing provider), *Type of Procedure, Type of Expense,* and *Expense Aggregator* (4 fields with different levels of detail), among others related to the Insurance charges, explanatory fields for the claim status, customer's coverage details, diagnosis, insured person personal details, the number of units (for the cases where the same procedures were done multiple times in the same episode, for example, if the patient had to undergo two x-rays, one on the right leg and one on the left), type of service, type of provider (in-network or out-of-network provider), and invoice's data. Table 4 exemplifies some entries of this table. As it can be seen for the same claim there can be several lines (if in the same appointment, the patient had to undertake several procedures), with the indication, when applicable, of the provider who prescribed a certain procedure, and the provider who executed it, as well as the type of service, and the field of medicine where it is inserted. Important to highlight that regarding the executing provider, there was an additional field: the executing provider at the health professional level (more detailed, in fact), which in this case was of no relevance – explaining by example, supposing that one patient has to perform an *MCDT* prescribed by the physician *X* and goes to clinic *A* to execute it, where the health professional *Y* was responsible for writing the final service report. For the project, in terms of relationship, what matters is the connection between physician *X* and Clinic *A*, whereas actor *Y* is not important - it is just an individual that was in some sense randomly assigned to this service - the patient did not choose it, the patient chose the clinic.

Within the project, the main purpose of this table was to retrieve the relationships between the provider who executed a certain procedure and the provider that prescribed it. However, that is not a direct task solvable with a simple query, several requirements had to be respected, mainly the type of procedure to consider, the medical specialties to comprise, the level of aggregation to follow, and what makes sense for the business to include or not – given that this is a complex table, most of its information was not used in the project, so it needs to be dealt with carefully.

Table 4 - Example for dataset (i)

Claim ID	Claim Line	Client ID	Claim Status	Procedure Start Date	Policy Start Date	Policy End Date	…	Provider ID (executer)	Provider ID (requester)	Total Value	Expense Aggregator 1
101	1	123	02	2017-01-07	2016-01-01	9999-12-31	…	4321	*NaN*	20	Consulta
101	2	123	02	2017-01-07	2016-01-01	9999-12-31	…	4321	2222	22	*MCDTs*
102	1	888	99	2019-08-14	2015-01-01	9999-12-31	…	2222	4444	89	Fármacos

To complement, get a better understanding of the business, and explain several approaches taken, Table 5 describes this database, mainly, the data format, dataset size, and key attributes (given the extent of information it contains, the focus will be on the most relevant fields), among others.

Table 5 – Data description report for dataset (i)

Dataset Name	*DMNL_SINISTROS*
Small Description	Médis Health insurance Claims/Billing data
Dataset Size	>10 Gb
Dimensions	Width: 130 variables, Height: 10 million rows
Variable Types	Claim ID (Int), Claim Line (Int), Client ID (Int), Claim Status (Char), Procedure and Policy Dates (Date Stamp), Provider ID - executer and requester (Int), Total Value (Float), Expense Aggregators (Str), etc.
Related Tasks	Filter *MCDTs*, Filter dates, Aggregate Claims, link to the table of providers' information, retrieve relationships, among others

Dataset *DMNL_PRESTADOR* (ii):

This table provides intel on the different providers inside the network of providers (providers with agreements and coverages), out-of-network providers are not included. It informs on *Provider ID, Provider Medical License Number, Provider Type, Provider Type Aggregator, Economic Group, Specialty* (there are two fields related to it, for the cases of providers with more than one specialty), and geographical details. The purpose of this table was to link more information to the actors of the network. Table 6 draws a sample of how the data is organized in this source table.

Table 6 - Example for dataset (ii)

Provider ID	Provider Name	Medical License	Provider Type	Economic Group	Specialty 1	Specialty 2	…	District	Zip code
00001001	Dr. Example (i)	12345	Médico	Not applicable	Cardiologia	Medicina Interna	…	Lisboa	1070-312

00001003	Clinic Example (ii)	54321	Consultas Ambulatório	Group example (i)	Terapia da Fala	NaN	…	Leiria	2400-082

Dataset *DMNL_AGREGADORES_DESPESA* (iii):

This dataset informs on the category of each type of procedure, according to different levels of aggregation, that is, how it is categorized on a more general level and a more detailed level. Table 7 is an exemplification of this data source. This table was not linked to any other table (as this information was also contained in the claims data table). However, its main purpose was to give some context to the business, understanding the different types of expense there can be and how they were allocated inside each level of the expense aggregator.

Table 7 - Example for dataset (iii)

Procedure	Expense Aggregator 1	Expense Aggregator 2	Expense Aggregator 3	Expense Aggregator 4
Fisioterapia	*MCDTs*	Medicina Física e Reabilitação	Tratamentos MFR	Exames e Tratamentos de Fisioterapia
SNS - Análises Clínicas	*MCDTs*	Patologia Clínica	Diversos	Análises Clínicas

Others:

Other used data sources include a table given by the clinical team with all the procedures of internment properly classified, a table with additional information on the episodes that were surgeries, a table identifying the pre-authorized procedures (provided additional information on certain claims), and several tables with medical billing data.

3.3. DATA PREPARATION

As it is common knowledge among data scientists, data preparation is a crucial step of the data science project. Literature often refers to it as the "80/20 rule", which as stated by (Snyder, 2019) means that 80% of the work comprises the collection, cleaning, and processing of data, with only the rest dedicated to the actual data analysis. The previous statement proved itself true. On one hand, by enabling the generation of a smaller dataset, it was possible to retain only the relevant information, remove errors and incoherence, and simultaneously have a quality dataset that could improve the efficiency and effectiveness of the applied techniques. On the other hand, through this process, several new findings on the data were found. Next, the main manipulations of the data are presented, both supported by analytical arguments and business context. These will be interspersed with findings from the data exploration phase. As previously stated, the tool that enabled these procedures was the SAS Enterprise Guide, which through the creation of several *Process Flows* assisted in having an overview of the sequence of the different tasks and queries, and of each task itself.

Firstly, the claims table (metadata details on page 33) was filtered according to the different assumptions (presented in 1.1.4) – this became a complex task as will be seen. To select claims from

the year 2021, the *procedures' start* and *end* dates were filtered (the procedure had to start and end within this time window). Next, using the field *Expense Aggregator 1* only *MCDTs* were selected. Through a small exploration, it was detected, by looking at the fields related to the status of the claim, that it would not make sense to include records of claims that were not approved, considered invalid, or that were still pending validation or approval (*claim status* and *refusal motive* variables). Additionally, testing purposes' records were removed (e.g., records whose *provider ID* was equal to "123456789") along with all the entries that did not have a *total billing value* above zero (being this a business environment, the billed amount matters).

Secondly, surgery and hospitalization-related records had to be removed (although there was a Boolean variable that identified hospitalizations, it was not accurate nor enough) - the identification of these cases was made in a separate process flow and was also based on the claim's dataset (although additional ones were necessary). So, to start, the claims from the claim's dataset were aggregated based on their *ID* (recalling, there were several lines for the same claim), and through the creation of several Boolean variables, claims with any procedure with outside providers were removed as well as cases with value zero in the hospitalization flag. Afterward, other relevant indicators were created (Indicators of *MCDT*, of hospitalization with an overnight stay, and of hospitalization without an overnight stay) to help with validation and with the creation of hospitalization categories. Next, there were cases of claims being separated (because the surgeon was out of network and all the other procedures were inside of the network) but belonged to the same surgical episode – these cases had to be identified (through an episode's table) and removed. Additionally, some claims were from oncology, but were not properly classified as hospitalization – the correction and proper identification were made in two ways: by pharmaceutical (if specific pharmaceuticals existed in the table) and by claims (through specific chemotherapy consumables). Several steps later, it was possible to identify all the hospitalizations without surgery. Through an identical process, it was also possible to flag surgeries with an overnight stay. These cases were then removed from our base table (obtained in the previous task).

Thirdly, claims that did not have outpatient coverage were removed as well as covid-19 tests, this was a simple querying procedure on the variables *Coverage* and *Expense Aggregator 3*, respectively.

Fourth, during the data exploration, several cases of invalid providers were found – these had to be dealt with. To start, given that the goal was to study relationships between providers (the executer and requester), entries without an underlying medical prescription were removed (cases with missing value in the field *Provider ID (requester)*) – with this, 27.4% (#2 054 560) of the entries of the work's table were removed. Next, it was important to know the providers that participated in each claim, as such every entity that through its *ID* could not be connected to the providers' table (dataset (ii), more details on page 33) was considered as an invalid provider, and thus removed – this step was performed on both types of the provider (the one who prescribed and the one who executed) and it resulted in the removal of 6.5% of the entries (353 351 out of the 5 447 173 rows) which comprised 1 497 distinct providers from the total of 25 786 different providers that the table contained at this point. Important to comprehend that one additional reason that could explain why some rows had the field provider (requester) blank is that the doctor who prescribed it is outside the network (or the refund was out-of-network) even if the entity that performed it (or the procedure itself) is within. Also, to notice that a small preprocessing was done on the providers' dataset before this stage, to keep only the relevant information (*provider ID, name, description, economic group,* and *geographic details*, among others).

The fifth step consisted of the removal of records related to discontinued lines of business and consideration only of claims for whom providers belonged inside the Médis network of providers.

The sixth task comprised the identification of *MCDTs'* records associated with emergencies, and their removal. To complete it several considerations had to be made. Contextualizing the reason behind this step, when exploring this data, several cases of *MCDTs* provided within the context of an emergency but not disclosed as such (for being billed in different claims) were detected, so it was necessary to find a workaround to correctly flag them. The solution was to group the entries by claims that occurred on the same day (that is, had the same start and end dates), for the same patient (had the same *customer ID*) and were supplied by the same provider (same *executing provider ID*) – if within a certain grouping, there was at least one emergency, all the records related to that aggregation would be recognized as being the product of an emergency and thus removed from the work's main table. The cases that were correctly classified as emergencies (allowing the identification of the others) had under the field Expense aggregator 1 the value of "Serviços de Urgência" and had outpatient coverage.

Also, throughout the main process flow, several variables were created, either to help with certain actions and steps or to safeguard against future situations where these would be needed. One case was the variable *Unit Billed Amount*, defined as the total billed amount divided by the number of units within that row.

In the seventh step, the specialty of the requesting doctor was obtained. Important to remember that within the base table, there was information on the specialty, but this was referred to the executing provider, which was not required. For example, if a cardiologist and a radiologist prescribe medical tests, even if different ones, when they are carried out both can be classified as imaging, although they correspond to quite different cases. In fact, what is desired knowing is the prescribing doctors' specialty (contained in the providers' table). However, there were physicians with multiple specialties, which raised the question: which one to consider? To solve this issue two considerations were made. First, if the provider had a single specialty, there was no doubt, that would be the retrieved one. Secondly, if there were two specialties, a connection had to be made, from the consultation where the prescription was made to the claim where the procedure was carried out, to retrieve the desired information. However, this was not a direct task, as in the claims table, under the line where the *MCDT* was billed, there was only data regarding the doctor who prescribed it, not the appointment itself. Consequently, it was necessary to recover and query the original claims dataset to obtain consultations-related claims. Afterward, foreign keys were needed, and with this several questions arose: how to connect a given medical exam, for example, to the consultation, given that they don't occur on the same day, or that a specific patient may have had several appointments with the same doctor during a given time window? Is considering the most recent appointment, prior to the procedure, the appropriate choice? To approach this situation, firstly, for these cases, the table of consultations was merged with the work table through the *requesting provider* and *patient IDs* (obtaining a table of all the appointments between these entities), and afterward the most recent appointment, before the procedure, was selected. From here, it was possible to choose the specialty for each case, as within the context of a consultation, although also coming from the claims table, the executing provider (for which there is specialty information) is the doctor who later prescribed a given procedure. Important to notice that there were some cases of patients who had more than one medical appointment with the same physician on the same day. However, this did not become an issue, as the corresponding specialty was the same, which was the required field.

Next, the eighth task involved the elimination of missing values, specifically, all the rows with missing values, in the fields of *specialty* and of *customer ID*, were removed – this approach was chosen, as on one hand, imputation of missing values would not produce one hundred percent reliable results potentially creating some bias in the network relationships, and on the other hand, at this stage, the dataset was still large and this created a substantiated way of resizing it to a slightly more manageable dimension.

To sum up, up to this point, there were still several lines for the same claims. Although, during all the data transformation stages, the goal of building a network was always present, and behind the several choices made throughout the process, it was important to think about possible levels of aggregations for this data – which level of detail was intended to show in the relationships? Providers linked by common patients with the weight being the total number of lines (procedures) the different claims comprised in the relationship contained? Or a weight equal to the number of claims? Or no weight at all? At that time, how the relationships would be built was not absolutely answered for two main reasons: each option would have different levels of complexity during the modeling, and being this an iterative process, as long as all the considerations were well made, the transformation from a more disaggregated dataset to a more aggregated one would be relatively easy, however, regardless of the choice, and also as part of the data exploration, several *MCDTs* contexts had to be analyzed, to understand if their behavior was consistent across the different groups of procedures, types of service, and expense aggregators.

Consequently, in the ninth step, and after consulting with the operational team, every entry that contained procedures related to clinical analysis had to be revised. That is, usually, for example, when a patient does blood tests, the samples are analyzed for several indicators, let's say ten, and in most cases, their unitary price is the same, but upon billing, for the same claim, ten equal entries would appear, or in some cases, they would be billed in ten separated claims (even if performed at once/same visit). It was concluded that, for these cases, it would make sense to aggregate the different claims related to the same episode into one and combine the different lines of the same claim into a single row, assigning to the variable *Number of Units*, the value corresponding to the number of aggregated items. However, regarding the clinical analysis, it was necessary to understand what the best variables were or set of conditions to identify them, as in the dataset several fields could be used, not identifying the same cases though. To explore the subject, several *group by* and counts were performed using combinations of the following variables: *expense aggregators* (from 1 to 4), *type of service*, and *procedure description*, aiming to identify which variables contained details that the others did not have, and jointly with the operations area, catch every clinical analysis-related case. As a result, the set of conditions that allowed identifying these occurrences was found, specifically, the *Expense Aggregator 3* had to be "Análises Clínicas", or it had to belong to the set {"Bioquímica, "Hematologia"," Imunologia", among others} and simultaneously assume the value "Patologia Clínica" in the *Expense Aggregator 2*, among other conditions imposed on the variables *Procedure description* and *Expense Aggregator 4*. Afterward, it was necessary to validate how to fill the different variables' fields upon aggregation. So, for the *procedure start* and *end* date values, it was considered the fields' minimum and maximum dates of the comprised items, respectively. For the *Service type*, within the same claim, only some cases had different values, thus being considered the mode. For the *Total value*, the aggregation function was the sum, recomputing consequently the *Unitary value* variable. For the remainder variables (as *customer ID*, providers-related fields, *line of business*, and *specialty*, among others) the values were equal among groups, so there were no issues. Additionally, some control

variables were added to the dataset regarding the original details of the grouped clinical analysis-related claims.

Next, in the tenth step, for the sake of consistency, all contexts billed similarly to clinical analysis had to be identified - it would make no sense to have mixes between aggregated and non-aggregated items. Therefore, together with both the team responsible for the billing process and the company's clinical team, two other cases were identified, physical therapy and Imagiology. Specifically, within physical therapy, there are cases of closed packages of sessions, meaning that one unit of this item comprises several sessions for a given treatment, which came into conflict with connections that existed to be repeated between the same set of providers, encompassing the same patient, for the different sessions that the patient performed. The approach here was to identify all the physical therapy cases (through the table of expense aggregators, dataset (iii), page 34), and then aggregate all cases of sessions referred to the same treatment as if they were a closed package. This process was like the clinical analysis one, aggregation-wise. Following, in the case of Imagiology, if for example, the requesting doctor would prescribe several exams to determine the source of a certain symptom, these as being part of the same plan would make sense to be billed together in the same claim, as such after identifying all the cases (through The *Expense Aggregator 3*) it was proceeded as previously.

It is important to highlight that these aggregations were made to have a more precise dataset. Though, if ultimately it would be decided at the network level to consider only if a given relationship between providers occurred or not – which in one of the modeling approaches turned out to happen - these last two steps would become unnecessary – but being this project based on an iterative methodology, it was preferable to have the option, even for testing purposes.

The eleventh step came as needed later in the project, as the requesting provider itself could work at the same place where the *MCDT* was performed, the requesting provider's associated entities were retrieved to further analyze them in subsequent steps. Specifically, this field would be beneficial to disclose the cases where people who had a consultation at a given healthcare facility with a certain physician (requesting doctor), had their prescribed *MCDT* executed at the same facility (executing provider). In terms of data dataset, it was created a Boolean variable, pointing out if for a given entry the provider (requester) had patronal connections to the healthcare facility where the service was performed.

The data transformation using *SAS EG,* ended with one additional feature selection, a newly added variable (zip codes for both types of providers, as it could be helpful when visualizing the network), and an aggregation of the different rows by claim ID. The fields that were kept were the following: *Client ID*, *Claim ID*, *Provider ID* (executer (*exec.*)), *Provider ID* (requester (*req.*)), *Provider Name* (executer), *Provider Name* (requester), *Physician Specialty* (from the requesting provider), *District* (requesting provider), *District* (performing provider), *Zip Code* (requesting provider), *Zip Code* (performing provider), and *Total Billed Amount*. Table 8 describes the resulting dataset at this stage. In short, it contains one line for each claim, describing the relationship between two providers, one who prescribed one *MCDT* and the one that executed it.

From here, the data exploration becomes simpler, as the dataset has a more manageable size. However, as observed in the previously described process flow, most of the data transformation tasks already involved a great amount of data exploration to make supported decisions.

Table 8 - Example of the transformed dataset

Claim ID	Client ID	Provider ID (*req.*)	Provider ID (*exec.*)	Provider Name (*req.*)	Provider Name (*exec.*)	Physician Specialty	District (*req.*)	District (*exec.*)	Zip Code (*req.*)	Zip Code (*exec.*)	Total Value
1	12345	4321	1111	Dr. José	Clínica ABC	Ortopedia	Lisboa	Santarém	xxxx-yyy	wwww-lll	200
2	88888	2222	2221	Dr. João	Clínica DEF	Clínica Geral	Porto	Braga	zzzz-vvv	dddd-eee	89
3	99999	3210	9182	Dra. Joana	Hospital Sto. André	Cardiologia	Lisboa	Lisboa	aaaa-bbb	aaaa-ccc	1500

Having a transformed table, free of inconsistencies and according to the different assumptions, the data transformation continued in *Python*, followed by additional data exploration.

Firstly, it was checked again if there were any missing values (the occurrences that were removed before did not comprehend every attribute). Some *NaN* values were found in the variables *District* (*req.*) and *Specialty* (*req.*), thus the corresponding rows were removed. Here, the approach was not an imputation of missing values, as these cases were residual (1 233 rows in the first case, 5 800 in the second, out of the 764 630 total rows).

One important dimension of the project was Physician *Specialty (req.).* As advanced in the Introduction of this thesis, the focus was Cardiology, however, before approaching the data itself the idea was to either analyze different networks according to the different specialties or to analyze the relationships of providers regardless of their specialty, using this attribute in a later step as a dimension to describe the results and to further conclude. At this point, the data comprised 56 different specialties across 757.597 different claims (Appendix C displays this distribution for the ten more frequent specialties). Also, by grouping the claims by physicians, there are several that prescribe *MCDTs* under different medical specialties (one doctor can have two specialties, as noticed previously).

Other relevant statistics, before resuming the dataset preparation, included: 21 715 different requesting doctors, 1 092 different executing providers, and 84 594 distinct connections between physicians and *MCDTs* providers. These values highlight the extension of the network underlying this project, which was an additional argument to follow the contingency plan, approaching only cardiology. But, before entering those details, there was one action left to build the network's adjacency matrix, which ultimately is one of the main objects under graph theory, as shown in the literature review of this document. It was necessary to group equal pairs of connections between providers and associate each connection with a weight given by the number of shared claims and the number of shared patients – ultimately only the number of shared claims was considered, however, both fields were kept until a certain point given that in some cases these values could be different (if one patient has more than one claim containing the same two providers), and for control and validation reasons. Figure 3.1 exemplifies the aggregation done.

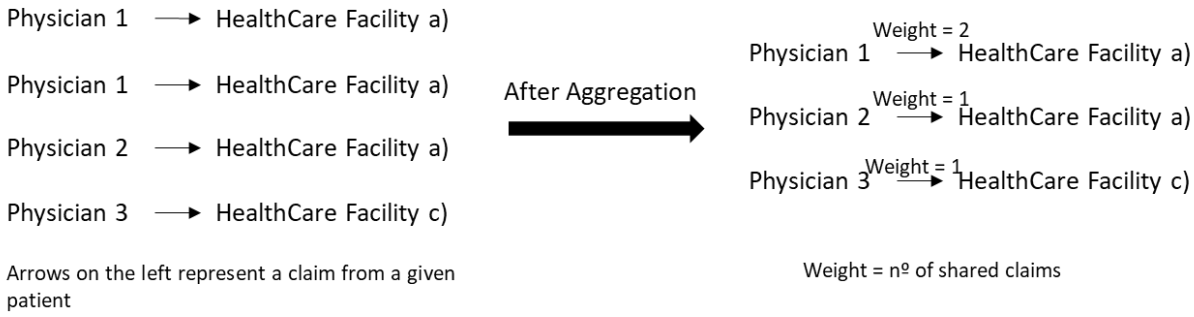

Figure 3.1 - Aggregation and weighting of equal pairs of connections (example)

In sum, it was obtained a table of unique connections between physicians that prescribe *MCDTs* and the healthcare facilities where these were performed, each of those relationships described by the number of shared patients, number of shared claims, and total billed amount. Each provider is characterized by their name and ID, and their geographic details (district and zip code).

The last step before entering the Modeling part of this project was to filter the table of connections to obtain only those within a cardiology context.

Overviewing the data, now focusing on the new universe, it included: 667 requesting providers, 455 executing providers, 3 175 distinct connections between physicians and *MCDTs* providers, 20 852 total claims, and 1 519 891.92 € total of the invoiced value.

3.4. MODELING AND RESULTS

Having prepared the raw data into quality data, the main instrument to proceed with the knowledge discovery is obtained. As such, during this section, all the main steps, to get as much expertise as possible about providers and their relationships, will be described.

3.4.1. Modeling Overview

To find the most relevant providers and connections between providers that request *MCDTs* and Healthcare Service providers where these are performed, this project had to contemplate several different tests and approaches. As explained and reasoned in Project Plan, SNA was the chosen examination process.

So, after extensive research on the topic under review, it was known a priori that the volume of data could develop into a problem. The calculation of the betweenness centrality would be one example of measure where the computational resources could become a liability, as such this was an important test to run given the 84 594 distinct connections across all specialties and the great importance centrality measures had for the scope of the project. The design of these first tests was as follows:

1. Choose critical points to evaluate their viability. In this case, graph projection, the four main centrality measures, mainly, degree, eigenvector, closeness, and betweenness centralities (for the two-mode projection of the network), and the drawing of the graph.
2. To evaluate the performance, the main *KPI* was the execution time of the previous items, understanding if it would be viable or not to study all the connections, or if the universe under analysis would have to be constrained.

3. Run the test for the whole graph, for each medical specialty graph individually, and the graph that included several different combinations of filters (thresholds-based).
4. Save the centrality measures' results as *.pickle* files in case these would be helpful later.

After following this plan, the conclusion was simple: a universe reduction had to be done. Analyzing the results (see Appendix D) together with a vision of the business and functioning of medicine, there is coherence. In General Family Medicine (classified as unfeasible), the physicians request the most various types of *MCDTs* given that usually people search them for multiple different reasons and symptomatology, while in a medical specialty like Cardiology, the domain is more constrained in types of *MCDTs*, and thus the group of healthcare service providers is not as diverse, making computational complexity of the different SNA algorithms more accessible.

Accordingly, initially, there were here two possible approaches, either follow the contingency plan (focus on cardiology) or along with the business team define a threshold to differentiate between relevant and non-relevant relationships, reducing the network size (for example, based on the number of the shared claims or patients between pairs of nodes). However, as pointed out earlier, the path to follow was the first, as the second did not prove itself as a feasible alternative. This means that after this experiment the analysis of these relationships became exclusive to the specialty of cardiology.

Having finalized this procedure, the plan was to build a two-mode network and make a dual projection, one on each bipartition of nodes. And, across all the steps, from network statistical measures (at both network and node levels) to community detection, conciliate the different methods to address both unipartite and bipartite networks, obtaining more robust conclusions, and somehow overcoming the limitations of each of the options, on one hand, the loss of information in the projection and on the other hand the limitation in the methods capable of approaching bipartite networks reliably.

3.4.2. Network construction

The process of building the main graph was made through Python, using the *Networkx* library, which is a module with elements capable of analyzing complex network structures, comprising several algorithms, analytical measures, and data structures (*NetworkX — NetworkX Documentation*, n.d.).

Before proceeding it was important to properly define and conceptualize the graph and its elements. From the start, it was known that the type of graph comprised by this SNA experiment was Bipartite. As underlined in the literature review of this thesis, specifically in Bipartite Networks, these structures, G=(V_1,V_2,E), are composed of three parts, two disjoint subsets (bipartitions) that make up the vertices set ($V(G) = V_1 \cup V_2$), and a set of edges E that represents the links between the nodes. In this study, respectively, one of the bipartitions is the set of providers *req.* and the other is the set of providers *exec.*, being these connect if they shared any claim or patient. Also, the representation of this network was made through an adjacency matrix (bi-adjacency matrix). The different symbols and notations used can be found in Table 9. Regarding one important relationship attribute, the weight, if introduced this would become a weighted bipartite network, reflecting on the bi-adjacency matrix which instead of being composed of *1s* and *0s* (in case there is a link or not), each entry would be a number corresponding to the weight of the associated link.

Table 9 – Graph Notations considered

Symbol	Description
$G(V_1, V_2, E)$	Providers bipartite graph
V	Vertices set, all the providers regardless of their type
V_1	Providers (*req.*) vertices set, only the physicians that prescribe *MCDTs,* $\{v_1^1, v_2^1, \dots, v_{n_1}^1\}$
V_2	Providers (*exec.*) vertices set, only the providers that perform *MCDTs,* $\{v_1^2, v_2^2, \dots, v_{n_2}^2\}$
n	Order of the graph G, the total number of providers = $\lvert V \rvert$
n_1	Total number of providers *req.* = $\lvert V_1 \rvert$
n_2	Total number of providers *exec.* = $\lvert V_2 \rvert$
B	Bi-adjacency Matrix for G
$G_1(V_1, E')$	Two-mode projection of G, on the providers' *req.* perspective
$G_2(V_2, E')$	Two-mode projection of G, on the providers' *exec.* perspective
A_1	Adjacency Matrix for G_1
A_2	Adjacency Matrix for G_2

Upon construction of the network, the initial step was adding V_1 (reflected as a list of the distinct providers (*req.*)) as the first set of nodes and V_2 as the second. Providing the binary bi-adjacency matrix, the edges were added to the structure $G(V_1, V_2, E)$, ending up with an unweighted bipartite network.

Furthermore, in a python environment, it was possible to add attributes to both edges and nodes, which could be helpful when constraining the study to a given group of individuals based on their attributes or their relationships. Specifically, the attributes contemplated were the frequency, the total billed amount, the number of shared patients, and the number of shared claims for the edges (attribute introduced with the weight nomenclature), and for the nodes, the provider's name, district, the vertices' set of origin (0 or 1), and the degree of the given node (added later). These elements enabled considering two versions of the graph, the unweighted bipartite graph and the weighted one, which was seen as an added value, by allowing one to choose which one to approach throughout the project. Figure 3.2 exemplifies the unweighted (a) and weighted (b) views of the bipartite network.

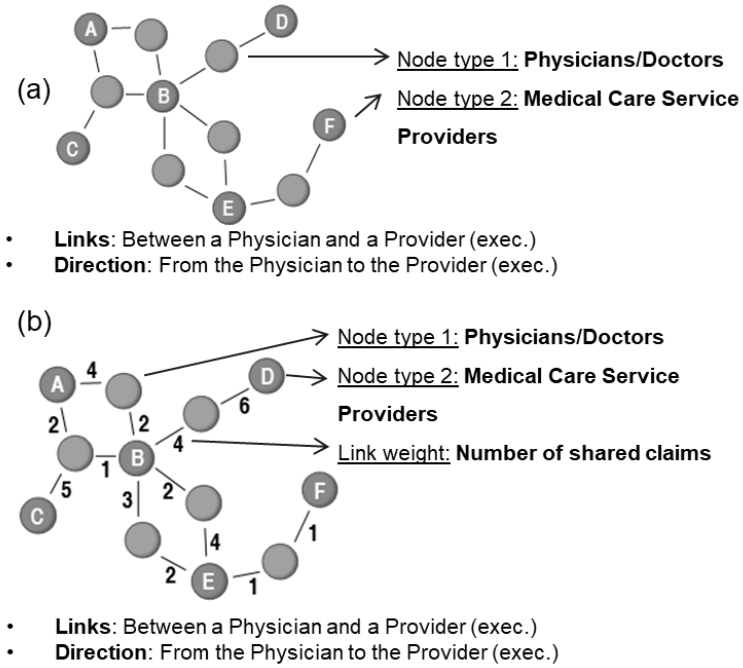

Figure 3.2 – Graph as an unweighted bipartite graph (a) and as a weighted one (b)

From all the data exploration and data transformation steps, it was clear that G was bipartite. Nonetheless, to confirm it mathematically, the spectral bipartivity was measured, resulting in a value of 1, which confirmed that the network was a representation of a bipartite graph.

3.4.3. First network insights

Before entering the specifics of this Social Network analysis with the study of measures and graph algorithms, it was important to do some initial exploration which, added to the previously acquired knowledge of the underlying data, could build a solid foundation for the upcoming steps.

The geographical dimension is an important subject within the context of this project. It is obvious that although this company operates in most of the national territory, the distribution of patients, providers, and market share varies between the different places. On one hand, the analysis of this aspect could help understand where there was a shortage of providers if there were frequent connections between providers from different districts. For instance, if there are a lot of people having a consultation with a cardiologist in Coimbra and then performing the *MCDT* in *Bragança*, that could indicate that the insurance company lacks agreements with cardiologists in *Bragança*, or that these do not fulfill the customers' needs. On the other hand, it could allow aesthetically more pleasing visualizations by providing positions for the nodes in the drawn plots. Additionally, by considering this element, it was possible to comprehend how distance affected the social phenomena that lead to the formation of this network, and how were the relationships created. According to (Scellato et al., 2010) several prior studies have examined how geographical distance impacts social relationships.

Accordingly, various values were analyzed per district of the referring physician, mainly, the distributions of requesting physicians, of executing facilities, and of the number of claims (Figure 3.3). By carefully examining the three distributions, it was possible to obtain some hints about the behavior of the underlying social entities. Most of the cardiology physicians involved in medical appointments with *MCDTs* being prescribed took place in *Lisboa* and *Porto*, however contrasting to the number of

providers in those same districts the distribution is distinct. Specifically, although most physicians are from Lisbon, the health service providers are more spread across other regions, giving some pointers on the possibility that people usually prefer to go to the capital and big medical centers to get checked up but end up performing the *MCDTs* in their own regions. This was an assumption that needed to be confirmed, as by looking at the number of claims, conclusions cannot be made.

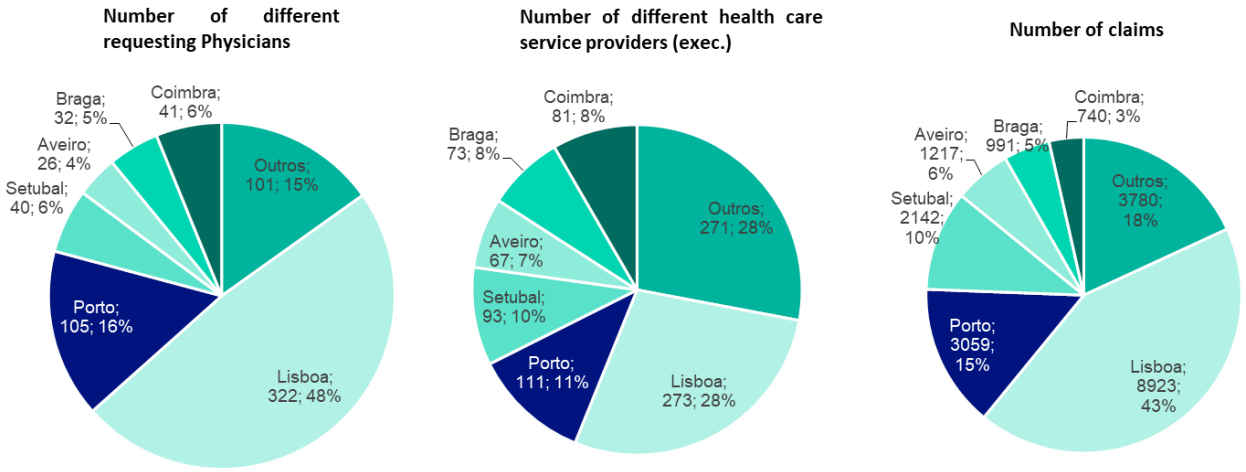

Figure 3.3 – Distributions per district of the requesting physician

Figure 3.4 displays the network, which just by analyzing it in plain sight made it hard to get detailed insights, however, it gave an overview of the network's structure, and it allowed measuring progress throughout this experiment, that is, by comparing the upcoming insights with this "threshold of knowledge" it was possible to know if there was any added value or not.

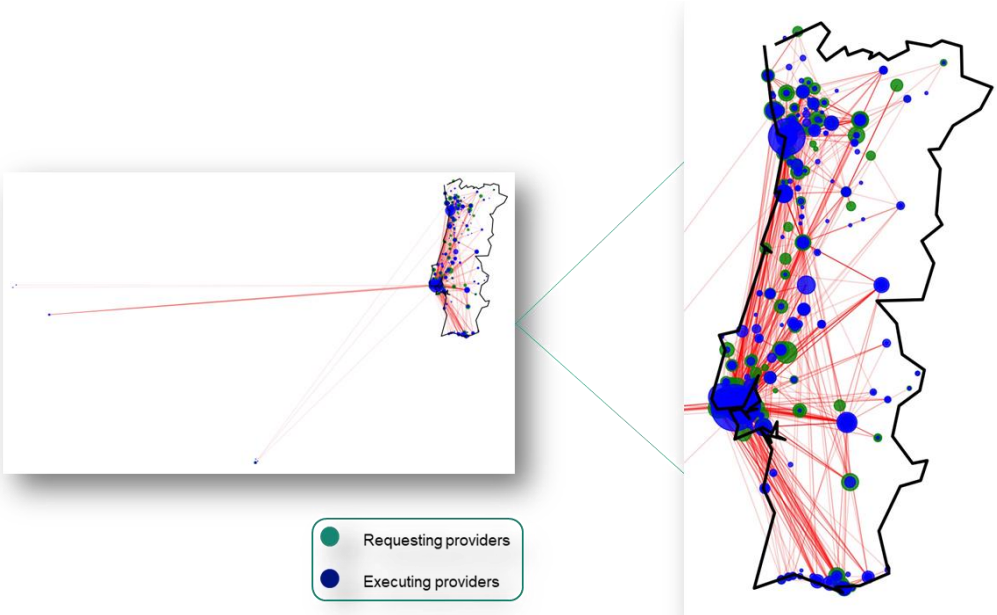

Figure 3.4 - Cardiology Network Visualization

3.4.4. Network Projection

As overviewed in section 2.1.2.2, two-mode projection is a topic of great relevance when approaching bipartite graphs. By converting the graph into a unipartite one, through the projection onto one of the

vertices partitions, most of the techniques that are specific for this type of graph can be implemented. However, the biggest drawback of this methodology is the loss of information. With these two thoughts in mind, the approach followed was making a dual projection (Figure 3.5). This method, according to (Everett & Borgatti, 2013), retains more structural information than using a single projection and enables the application of methods that otherwise could not be applied to data with a bipartite underlying structure. The approach followed was a mix between the previously indicated method (but with a twist) and the network approach in its bipartite structure. In sum, the analysis of the providers' network was based on the following items:

1. Dual projection of the unweighted bipartite network into two unweighted unipartite networks (networks (b) and (c) in Figure 3.5)
    a. Network of physicians: projection of the providers' *req.* vertices side based on their connectivity to the providers *exec.*, that is, by shared health units
    b. Network of healthcare units: projection of the providers' *exec.* vertices side based on their connectivity to the providers *req.*, that is, by shared physicians
2. Dual projection of the unweighted bipartite network into two weighted unipartite networks whose weight equals the number of shared neighbors
3. Dual projection of the unweighted bipartite network into two unipartite multigraphs whose multiple edges represent the various shared neighbors – done for experimental reasons, not being taken forward. It added a degree of computational complexity to the network (the amount of information stored was greater), with information which in a way could be retained by the weighted version of the network
4. Unweighted Bipartite Network (direct analysis)
5. Weighted Bipartite Network – due to the scarcity of algorithms and the research gap in this type of structure topic, this network ended up not being used as an additional layer of analysis

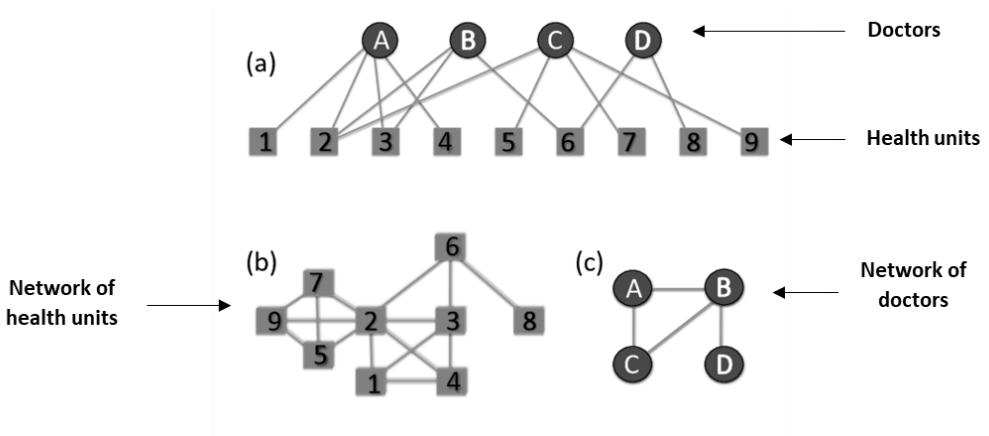

Figure 3.5 – Dual projection of the network of providers

Specifically, in the python environment, these conversions were made recurring to the available bipartite algorithms under the *networkx* package.

After carrying out these steps and with different analysis options available, resulting from the different perspectives of the network obtained, it was possible to proceed with a statistical analysis of the network, explained in detail in the next section of this work.

3.4.5. Network Statistical Measures

As known, statistical analysis is an important element in the exploration of the different actors in the network, and their connections too. Some of the key network statistics, measures, algorithms, and other metrics were applied, both at the actor level (individual level) and the network level.

To proceed with the study of the network at the individual level, the first tool used were centrality measures. The goal was to determine the importance of each node from different perspectives, as each of these measures has its definition. Starting by calculating these values on the unweighted unipartite networks (the result of the two two-mode projections performed), Table 10 – Centrality Measures Definitions, explains the different meanings of each.

Table 10 – Centrality Measures Definitions

Centrality Measure	**Definition in the physicians' network**	**Definition in the providers' (*exec.*) network**	**Measure Summary**
Degree Centrality	The doctor's importance is correlated with the number of healthcare units they share with other doctors, that is, the more providers *exec.* they share, the more important the doctor is thought to be.	The healthcare service providers' (provider *exec.*) importance is correlated with the number of physicians shared with other healthcare service providers, that is, the more providers *req.* they share, the more important the unit is thought to be.	Finds very connected actors
Closeness Centrality	The physicians are ranked by proximity/closeness (number of connections between doctors) to the other physicians in the network.	The providers *exec.* are ranked by proximity/closeness (number of connections between providers *exec.*) to the other providers *exec.* in the network.	Finds actors that are good network broadcasters. High values mean a higher likelihood of communication
Eigenvector Centrality	The importance of a given physician is linked to the importance of the physicians related to it, that is, the more healthcare units a doctor has in common with other prominent doctors, the more significant they are, and vice-versa	The importance of a healthcare unit is linked to the importance of the facilities related to it, that is, the more physicians a healthcare unit has in common with other prominent units, the more significant they are, and vice-versa	Finds actors that are influential across the network (not just at the neighborhood level)
Betweenness Centrality	The influential physicians are the central ones, that is, the greater their chance of connecting with other doctors, the more influential/centered they are	The influential healthcare service providers are the central ones, that is, the greater their chance of connecting with other units, the more influential/centered they are	Finds actors that act like bridges (influencers) in the flow of the network

Important to highlight that the underlying idea of these measures was to understand who the providers (requesters and executers) with a larger effect in the network of this health insurance

company were, under the domain of *MCDTs* cardiology-related. Thus, if the most influential/central providers of the network stop collaborating with Médis, other providers might also be persuaded to leave the network. On the other side, if the company concludes that the cost-benefit of keeping agreements with low influential providers does not justify, some actions might be taken.

Having conceptualized the previous notions, it was time to put them into practice. As a result, this segment started with what some intellectuals affirm as being the simplest measure of centrality, the degree centrality. From the direct analysis of the degree centrality (for the Bipartite graph as a whole), and through the *networkx* module bipartite functions it was possible to calculate this value for both weighted and unweighted versions. Thus, by assessing the number of connections each node has with other nodes in the network, the results were as follows. Important to notice that throughout this thesis, the providers' real names were anonymized for privacy reasons.

Table 11 – Providers *req.* Degree centrality for the unweighted and weighted bipartite graph (Top 4)

Provider *req.* Name	**Degree Centrality (unweighted bipartite graph)**
Dr. LB	25
Dr. VPM	22
Dr. LMDO	21
Dr. JESL	19
Provider req. Name	**Degree Centrality (weighted bipartite graph)**
Dr. RNAS	620
Dr. JG	401
Dr. VPM	370
Dr. VG	368

Table 12 - Providers *exec.* Degree centrality for the unweighted and weighted bipartite graph (Top 4)

Provider *exec.* Name	**Degree Centrality (unweighted bipartite graph)**
Lab Dr. JCAC	206
Lab Dr. CST	129
HCHF	121
SIi	118
Provider *exec.* Name	**Degree Centrality (weighted bipartite graph)**
Lab Dr. JCAC	1139
CSM	980
Clin CS	742
Lab Dr. CST	668

As it is possible to observe through Table 11 and Table 12, the ranking varies if using the weight or not, which indicates that the conclusions will not be the same for the two graphs. Nonetheless, although the positions in the ranking are not the same, there was consistency in considering a given node as important or not, which ultimately is more important. Significant to highlight that being the degree calculated based on the number of connections, evaluating the degree of the nodes from the bipartite graph or through the dual projections will provide equal results.

Also, by observing the nodes with lower degree values, some highlights could be made, mainly, *Dr. NRS,* and *Dr. JM*, among others for the side of the physicians, and *ASMJD*, and *SCSBE*, among others for the side of the health units. The company can use these indicators to trace a plan if it needs to shrink its network of providers.

Having calculated the degree for each node, this was added as a node attribute, in case it was necessary for filtration or analytical purposes in future steps. The nodes' degree provides an analysis at the individual level, which made it possible to differentiate between important/highly connected providers and less important ones. However, to have an overview of the graph structure, the examination of the degree distribution (frequency count for each degree's occurrence) provided valuable insights. Figure 3.6 presents the degree distribution for each of the vertices' partitions, or equivalently, for each of the unweighted unipartite projections of the original graph. Observing the distribution of the physicians it was possible to notice that the distribution is not as skewed as in the case of the units of healthcare service providers. Meaning that, while in the first case, there are quite some physicians with a moderate degree, in the second case it is clear that the high-degree cases are a minority compared to the ones with a lower degree – this type of distribution can be associated to a scale-free network, which as mentioned by Both (Foster et al., 2021; Newman, 2010) can lead to preferential attachment (new nodes that come into the network are more likely to link to high degree actors), in terms of business this can mean that the branch of business involving these individuals has the potential of growing so it is important to take the most advantage of that.

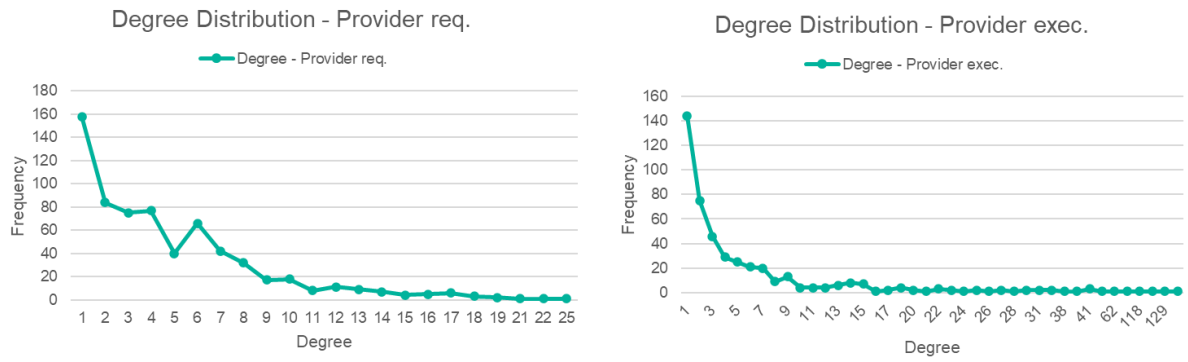

Figure 3.6 - Degree Distribution (of the unweighted bipartite network)

As the consideration or not of the weight (number of shared claims) was present throughout the project, it was important to understand its impact in the degree distribution of the network. In short, the goal was to compare the popularity's distribution (given by the importance, and consequently degree) with the strength's distribution of the several links (given by the weighted degree). To compute it, for each projection (providers *req.* and providers *exec.*) it was contrasted the relationship between frequency and degree of the simple graph to the one originated by the degree of the multigraph, which was the same as comparing the simple degree's distribution with the degree's distribution from the

weighted bipartite graph - Figure 3.7 displays this analysis. Observing the different values, although their order of magnitude is different, it is easy to conclude that the direction of variation is the same, meaning that the number of important individuals and less important ones is similar when looking at the weighted and the unweighted degree (in both nodes' sides).

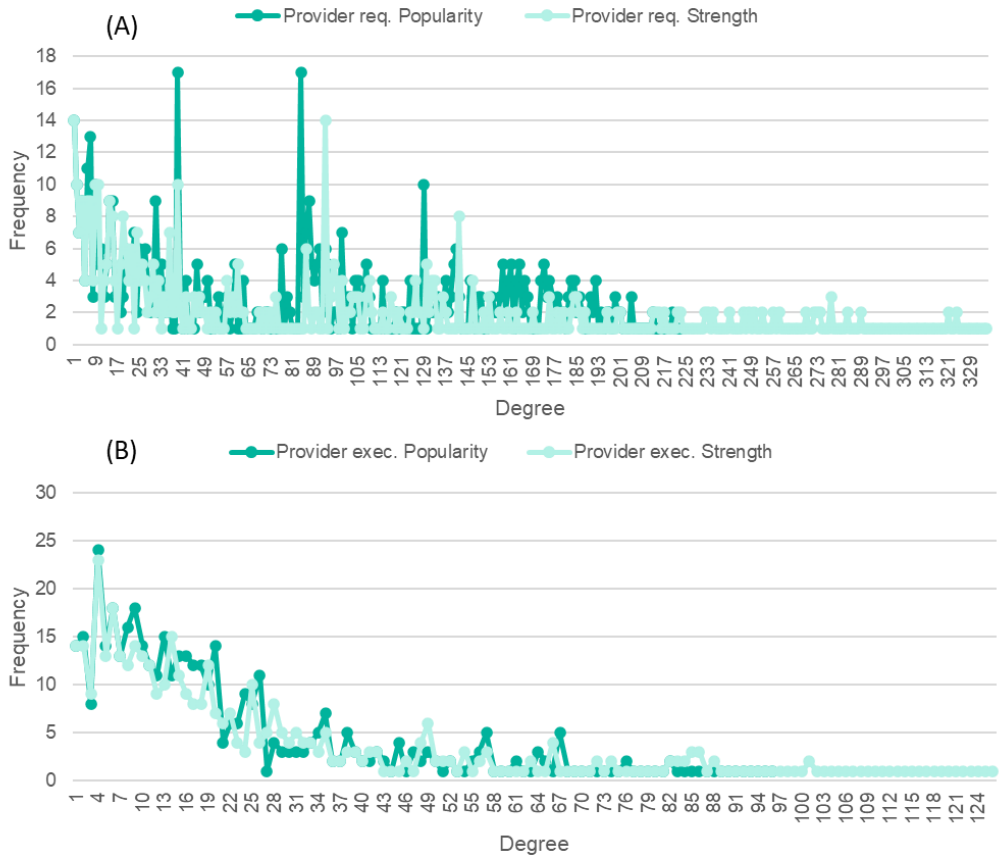

Figure 3.7 – Degree Distribution: Popularity and Strength, for the Providers *req.* (A) and Providers *exec.* (B)

To further analyze the relationship between popularity and strength, Figure 3.8 displays two scatter plots that evidence that for the different nodes, the importance determined by each of the indicators varies in the same direction, meaning that important individuals are classified as such, both looking at the weighted degree and the unweighted one and vice versa – this was verified for both nodes' sides.

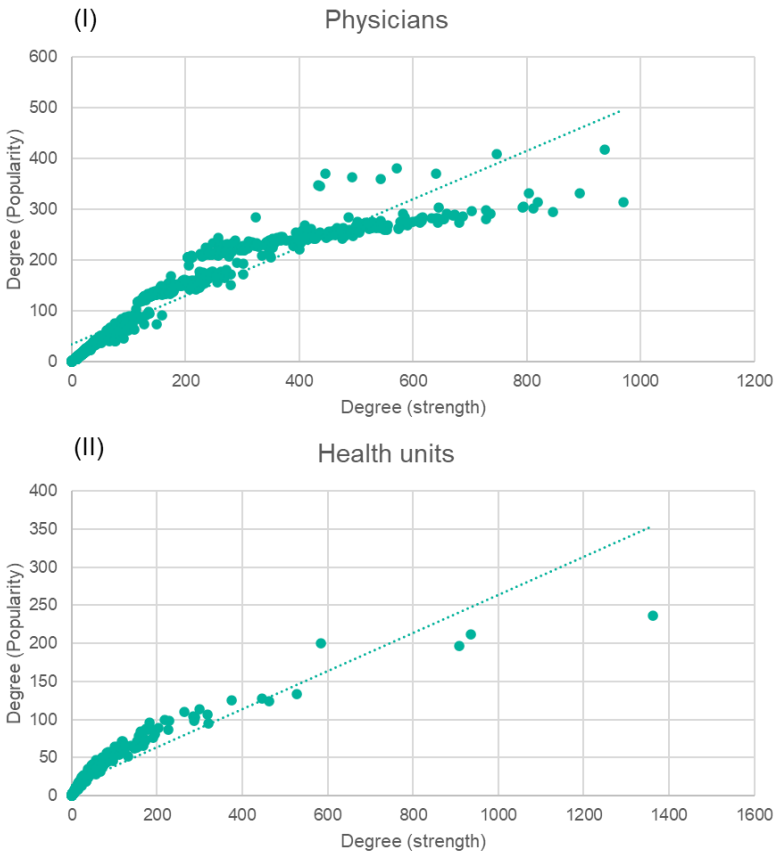

Figure 3.8 - Degree: Popularity versus Strength, for the Providers *req.* (I) and Providers *exec.* (II)

Resuming the analysis at the node level, after reviewing the degree distribution at the network level, the next centrality measure approached was the closeness centrality, which, as previously explained, is an indicator of the proximity between the several actors of the network.

At this point, the first limitations on the use of the weighted unipartite projections emerged (although there are some adaptations such as Dijkstra's algorithm, these were not implemented). So, taking the dual projection approach, this statistic was measured for both unweighted projections of the original network, the one on the physicians' side and the one on the healthcare service provider's side. It should be noted that it was chosen to use the normalized version of the closeness centrality.

Accordingly, Table 13 displays the four physicians who are more central (best broadcasters) in the network regarding this measure, whereas Table 14 displays the same but concerning the providers (*exec.*).

Table 13 - Providers *req.* normalized closeness centrality from the unweighted projection G_1 (Top 4)

Provider *req.* Name	**Closeness Centrality (unweighted projection G_1)**
Dr. NC	0.709
Dr. DT	0.702
Dr. PF	0.680
Dra. PA	0.673

Table 14 - Providers *exec.* normalized closeness centrality from the unweighted projection G_2 (Top 4)

Provider *exec.* Name	Closeness Centrality (unweighted projection G_2)
Lab Dr. JCAC	0.650
HCHF	0.619
Lab Dr. CST	0.616
SIi	0.598

Observing the case of the physicians, and comparing it with the degree centrality, it is direct that the ranking changed. In the case of the providers *exec.* the differences are not so notable, in fact, they even have some similarities.

Next, it was important to understand if these results were consistent in the direct analysis of the bipartite graph. As such, taking all the nodes regardless of their type, Table 15 shows the most important providers.

Table 15 – Direct Analysis of the bipartite graph: Top 6 nodes in closeness centrality

Provider Name	Provider Type	Closeness Centrality (bipartite graph G)
Dra. CV	Provider *req.*	2
CCC	Provider *exec.*	2
Dr. JRP	Provider *req.*	2
XC	Provider *exec.*	2
APSE	Provider *exec.*	2
EMA	Provider *exec.*	2

Comparing Table 15 – Direct Analysis of the bipartite graph: Top 6 nodes in closeness centrality with Table 13 - Providers *req.* normalized closeness centrality from the unweighted projection G_1 (Top 4) and Table 14 - Providers *exec.* normalized closeness centrality from the unweighted projection G_2 (Top 4), the conclusions are quite different when looking at the network as a whole or the projections individually. Ultimately, some decisions had to be made on how to reconcile results from several perspectives (approached later in this thesis).

Following with the Eigenvector Centrality, due to its limitations on the direct analysis, it was only measured for the dual projections (both unweighted and weighted). As studied before, this indicator computes the centrality of a given node based on its neighbors' centralities. Table 16 and Table 17 display this analysis, respectively for the providers *req.* and the providers *exec.* sides, in both unweighted and weighted unipartite projections, understanding if there are any variations on the ranking of importance mirrored by this statistic.

Table 16 - Providers *req. Eigencentrality* from the unweighted and weighted projections of G (Top 4)

Provider *req.* Name	Eigenvector Centrality (unweighted projection G_1)
Dr. NC	0.079
Dr. DT	0.077
Prof Dra. CAC	0.074
Dr. DNMAEM	0.074
Provider *req.* Name	**Eigenvector Centrality (weighted projection of G)**
Dr. LB	0,143
Dr. LMDO	0,127
Dr. JCCC	0,125
Dr. NC	0,122

Table 17 - Providers *exec. Eigencentrality* from the unweighted and weighted projections of G (Top 4)

Provider *exec.* Name	Eigenvector Centrality (unweighted projection G_2)
Lab Dr. JCAC	0.208
HCHF	0.196
SIi	0.192
CIiSS	0.157
Provider *exec.* Name	**Eigenvector Centrality (weighted projection of G)**
Lab Dr. JCAC	0,487
SIi	0,387
HCHF	0,383
CIiSS	0,247

Analyzing the previous results, the ranking, although not equal, was consistent with the one obtained from the closeness centrality measure. Also, when comparing the values from the unweighted projections and the weighted ones there are similarities, especially on the providers *exec.* side.

Finally, the last centrality measure overviewed was the betweenness centrality. The goal of evaluating it was to understand who were the providers that influenced the company's activity the most. In short, here the concept of centrality is associated with being influential. In this project, this measure showed itself as complete. On one side, the notion of referral can be associated with the influence of a given provider, on the other side it was possible to have several views of this indicator: on the dual projections, unweighted and weighted, and on the direct analysis of the bipartite graph. Also, to ease the comparations, it was used the normalized version of this indicator. Table 18 and Table 19 show the results for the dual projections, respectively for the providers *req.* side and for the providers *exec.* side.

Table 18 - Providers *req.* Betweenness centrality from the unweighted and weighted projections of G (Top 4)

Provider *req.* Name	Betweenness Centrality (unweighted projection G_1)
Dr. PF	0.044
Dra. PA	0.033
Dr. DT	0.031
Dr. NC	0.029
Provider *req.* Name	**Betweenness Centrality (weighted projection of G)**
Dra. PA	0.039
Dr. PF	0.033
Dr. JG	0.026
Dr. RS	0.026

Table 19 - Providers *exec.* Betweenness centrality from the unweighted and weighted projections of G (Top 4)

Provider *exec.* Name	Betweenness Centrality (unweighted projection G_2)
Lab Dr. CST	0.186
Lab Dr. JCAC	0.136
HCHF	0.095
Sli	0.054
Provider *exec.* Name	**Betweenness Centrality (weighted projection of G)**
Lab Dr. CST	0.109
HCHF	0.073
Lab Dr. JCAC	0.068
ALAC	0.048

Additionally, and like the closeness centrality, when it came to the direct analysis of the bipartite graph, there were available adaptations of betweenness centrality. However, only to the unweighted bipartite graph. Still, being able to calculate it was already an added value for the project. Table 20 displays the most influential nodes in the network (for both node types).

Table 20 - Direct Analysis of the bipartite graph: Top 6 nodes in betweenness centrality

Provider Name	Provider Type	Betweenness Centrality (bipartite graph G)
Lab Dr. JCAC	Provider *exec.*	0.271

Lab Dr. CST	Provider *exec.*	0.268
HCHF	Provider *exec.*	0.115
SIi	Provider *exec.*	0.080
Dr. PF	Provider *req.*	0.054
Dr. DT	Provider *req.*	0.039

In this last measure, there are similarities between the different perspectives, making it possible to identify the actors that influence the flow of the network the most.

However, when comparing the several centrality measures there are some differences, which raised the question: how to conclude the importance of the several providers? The answer ended up being simple – it was not a matter of choosing one approach or the other, these concepts have different notions of importance, so according to the context, it might be more correct to use the insights obtained from a certain indicator regarding the others. As such, for business purposes, the various conclusions were explained in each context, making it possible to know in each situation the one to point towards. Regardless, overviewing all the previous results, similarities in the ranking can also be identified - there are actors quite frequently classified as central, reinforcing, consequently, their role in the network. Nonetheless, Figure 3.9 allows an understanding of the relationships between the several centrality statistics, evidencing the previous aspects (the measures include degree, betweenness, and closeness centralities for the bipartite graph; the eigenvector centrality was not included due to its limitations when applied to this type of network, as mentioned previously).

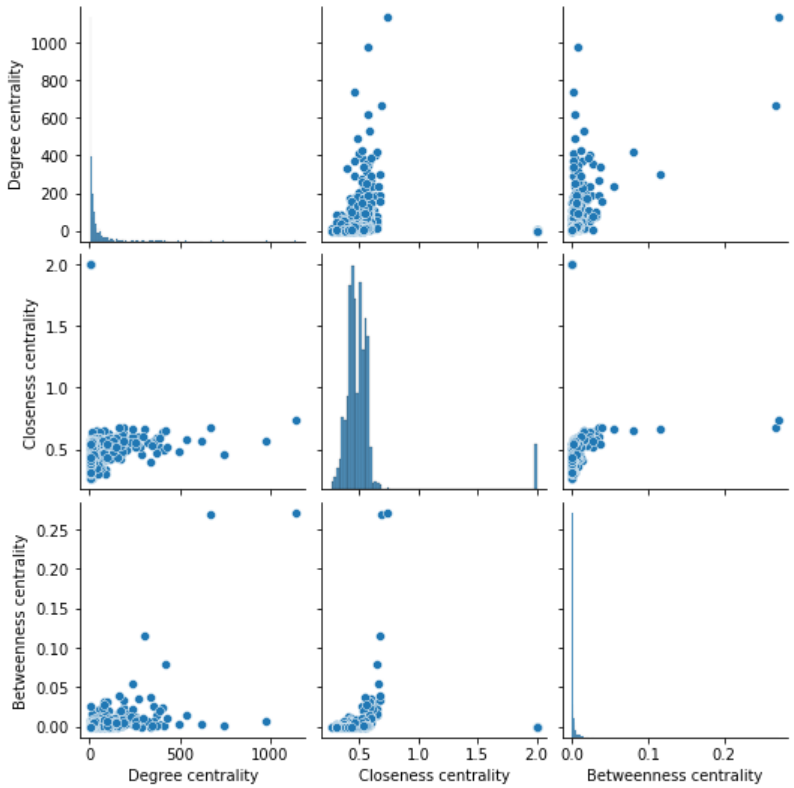

Figure 3.9 – Relationships between centralities (degree, closeness, and betweenness) of the bipartite graph G

Lastly, regarding the centralities, to have an overview of the distribution of these values across the several nodes of the network, the network was drawn using a centrality-based color map. Specifically, this was done for the projections G_1 and G_2, using the *Fruchterman-Reingold* layout. For G_1, especially on the measures of closeness and betweenness centrality, it is clear that using this layout the nodes with higher importance are positioned in the center of the visualization. Although for the projection G_2 of the providers *exec.* this behavior is not as evident, closeness centrality still stands out under this layout (see Appendix E).

Another insightful measure was the local clustering coefficient, which is a local measure of density that indicates the likelihood of the neighbors of a given node being connected to one another. As it is possible to observe in Table 21 from the top-ranked nodes in terms of clustering coefficient, there is no link to the previous indicators of importance. Here, having a high value means that locally the neighborhood of the node is well connected.

Table 21 - Local Clustering Coefficient for the bipartite graph G (Top 6)

Provider Name	Provider Type	Local Clustering Coefficient (bipartite graph G)
CMPV	Provider *exec.*	0.625
BEF	Provider *exec.*	0.625
Dr. NS	Provider *req.*	0.533
Dra. SG	Provider *req.*	0.533
Dr. VMM	Provider *req.*	0.472
Dr. NPM	Provider *req.*	0.472

Up until this point, the analysis was focused on the individuals that participate in the cardiology network of providers. Nevertheless, it is also valuable to understand the structure of the network. So, the next step was to move this statistical analysis to the network level. So far, it was examined the degree distribution, but there are additional metrics that can help understand the topology of the network.

Now, looking at the global bipartite clustering coefficient. This measure of density indicates how connected the neighbors of the several vertices are to one another, that is, the nodes' tendency to cluster together. In other words, it can be seen as the probability (on average) that any two neighbors of a node are also neighbors between themselves, or mathematically speaking, the average bipartite (local) clustering coefficient. In this project, this value amounted to 0.143, which is not a value of magnitude (differentiating on the set of nodes, the value obtained was 0.162 and 0.115 for the providers *req.* and providers *exec.* sides, respectively). To understand how dense the network of physicians was, and how well connected were the health units in their network, this measure was also calculated in relation to the projections of the network. For the physicians, the global clustering coefficient was 0.804 and 0.086 in the unweighted projection and weighted one, respectively. Whereas for the providers *exec.* it measured 0.675 and 0.012 (unweighted and weighted projections, respectively). These results indicate that the two unweighted unipartite networks that resulted from the two-mode projection have groups of nodes with a tendency to cluster. However, it is important to highlight that high values do not necessarily mean a general behavior across all elements in the

network and vice versa for lower values. By looking at Figure 3.10 it is possible to observe how diverse are the local clustering coefficients throughout the bipartite network. Although some nodes have higher values, there is still a great amount with lower values.

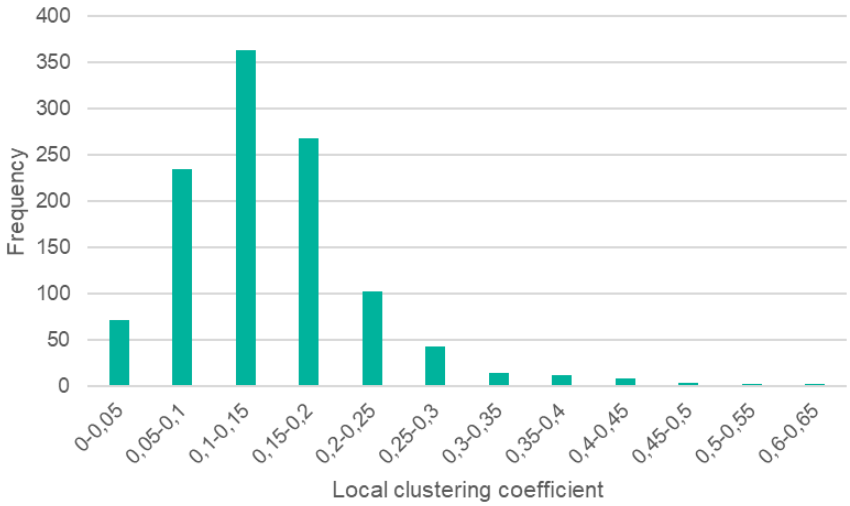

Figure 3.10 – Distribution of the Local clustering coefficient

Now, instead of taking the average of the adaptation of the classical definition of clustering coefficient for bipartite graphs, it was calculated the *Latapy* redefinition for this type of graph, obtaining an average value of 0.143 for the whole graph. Additionally, taking the *Robins and Alexander* bipartite clustering it was obtained 0.164. As can be seen, there were little differences between the different redefinitions of the classical approach. Furthermore, several quantitative indicators were measured to characterize the network in hands. Although the majority was previously mentioned, Table 22, displays the profile of the several networks utilized.

Table 22 – Graphs' Characteristics

Graph Characteristic	**Bipartite Graph G**	**Unweighted projection G_1**	**Unweighted projection G_2**	**Weighted projection (providers *req.*) of G**	**Weighted projection (providers *exec.*) of G**
Order (No. of Nodes)	1 122	667	455	667	455
Size (No. of Edges)	3 175	42 516	5 747	42 516	5 747
Average Degree	5.66 (nodes' degree values ranging in [1;206])	127.48 (nodes' degree values ranging in [0;417])	25.26 (nodes' degree values ranging in [0;236])	196.25 (nodes' degree values ranging in [0;970])	49.42 (nodes' degree values ranging in [0;1362])
Clustering Coefficient	0.143	0.804	0.675	0.086	0.012
Density	0.010	0.191	0.056	**	**
Diameter, radius, and center	**	*NaN* (the graph is not connected)	*NaN* (the graph is not connected)	*NaN* (the graph is not connected)	*NaN* (the graph is not connected)

Number of communities[1]	**	19	20	18	20

**Not applicable, or no implementations of the metric(s) available.

The observation of the average degree of previous results points to the average importance of the several nodes in the network. In this case, it was concluded that overall, in the mentioned networks, the values are low - despite having elements with extremely high values, there are others with lower values who contributed to the final value.

Also, the networks have low density. Focusing on the bipartite graph, it means that the network is sparse, implying that the information flow is not as easily done as in dense networks. Even though there is no formal definition of sparsity, this is a recurrent topic in large real-world networks, but with few literature advances on it.

3.4.6. Additional metrics

In addition to the previously presented insights, two metrics were created to have a better understanding of the several relationships – these will be elaborated on in this section.

Firstly, to understand the most important connections in the network, these were overviewed in terms of frequency. Being the network from a business context, the overall numbers matter, not just if a certain provider is relevant or not, as such the number of shared claims between pairs of actors was analyzed. From this assessment, several connections were identified as being more frequent than the average, examples included: i) link between *Dr. RNAS* and *Clin. CS* (556 shared claims), ii) link between *Dr. JG* and *CDCJG* (338 shared claims), iii) link between *Dr. VG* and *CDCVG* (278 shared claims), and iv) link between *Dr. VPM* and *C-Clin. CS* (274 shared claims), among others.

One of the goals of the connections' examination was to identify biased relationships or some tendencies. To approach it, the procedure outlined included:

1. The creation of measures to classify the networks' connections, mainly:
    a. *Ratio A* $= \frac{Total\ number\ of\ shared\ claims}{Total\ number\ of\ \boldsymbol{doctor's\ claims}}$, $for\ each\ connection$, which is equivalent to the proportion of this doctor's prescriptions performed at this facility, meaning that if the value tends to 0 then the number of shared claims assumes a small proportion of the physician's total number of claims, whereas values close to 1 mean that the volume of shared claims assumes a large proportion of the physician's total number of claims.

    b. *Ratio B* $= \frac{Total\ number\ of\ shared\ claims}{Total\ number\ of\ \boldsymbol{MCDTs\ provider's\ claims}}$, $for\ each\ connection$, which is equivalent to the proportion of prescriptions performed by this facility that were prescribed by the doctor in the relationship. Similarly, a ratio close to 0 means that the shared claims assume a small proportion of the *MCDTs'* provider's total number of claims, while values close to 1 mean that the shared claims assume a large proportion of the *MCDTs'* provider total number of claims.

[1] Based on the highest modularity partition generated by the Louvain algorithm. This topic will be overviewed in subsequent sections.

2. Accordingly, to find a tendency in the connections, among the ones that share a significant amount of claims, the process was:
    a. Filter out connections whose number of claims is not above the quantile ninety (in this case 12). This step was important as the company wanted to identify tendential connections among the ones that are relevant in the business spectrum.
    b. Determine the ones that have higher ratios.

Following this methodology, several potentially biased links between physicians and the healthcare facilities that perform their prescriptions were flagged. Examples include:

- The relationship between *Dr. FG* and *CCMME*, which had a ratio A of 1 (100% of this doctor's prescriptions were performed at this facility) and shared 61 claims, and the connection between *Dra. MVPT* and *DGDPAFPS* with a ratio A of 0.98 (98% of this doctor's prescriptions were performed at this facility) also share 61 claims.
- According to Ratio B, some connections flagged were between *Dr. JMDSES* and *PCT* with 76 shared claims and a ratio of 1 (100% of this facility's prescriptions are prescribed by this doctor), and between *Dr. LP* and *CCJ* with 58 shared claims and a ratio of 1 as well.

Nevertheless, after carefully analyzing the results, it was noticed that most of the detected connections are between a doctor who also works at the facility where the procedure (*MCDT*) was made, which is a natural relation. Also, the "big groups" (large-scale healthcare centers that provide their patients with the most diverse types of service, care, and doctors) have a great presence in these results. As an attempt of finding important relationships outside of these circumstances, two filters were imposed:

1. The health care service provider (*exec.*) cannot belong to the group of entities where the prescribing doctor works (information contained in the transformed dataset).
2. The health care service provider (*exec.*) cannot belong to the following health groups: *Luz, Cuf, Trofa, Lusiadas Saúde,* and *Hpa,* which in collaboration with the operational teams of the company were identified as members of the set of "big groups".

With this universe reduction, the network obtained went from having 3 175 connections, 667 doctors, and 455 executing providers to 59 distinct connections between physicians and *MCDTs* providers (reduction of 98%), 46 requesting doctors (reduction of 93%), and 20 executing providers (reduction of 95%). Following this new setting, it was performed an analysis of the distribution of the previous ratios (Figure 3.11), which gave an overview of the proportion of relationships potentially biased in this universe. Detailing on the relationship level, other connections flagged comprised:

- The relationship between *Dra. RMG* and *LAP*, which had a ratio A of 0.96 (96% of this doctor's prescriptions were performed at this facility) and shared 24 claims, and the connection between *Dra. MON* and *CC* with a ratio A of 0.90 (90% of this doctor's prescriptions were performed at this facility) and sharing 45 claims.
- According to Ratio B, some connections flagged were between *Dra. MON* and *CC* with 45 shared claims and a ratio B of 0.87 (87% of this facility's prescriptions were prescribed by this doctor), and, as in ratio A, the connection between *Dra. RMG* and *LAP* with 24 shared claims and a ratio B of 0.86 (86% of this facility's prescriptions are prescribed by this doctor).

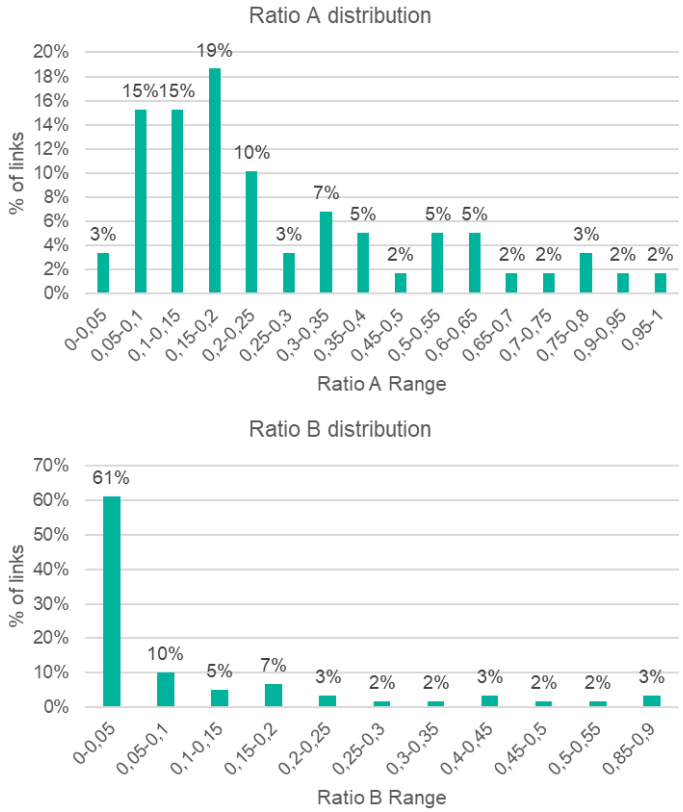

Figure 3.11 - Ratios' distributions on the network with restrictions on *big groups* and doctors' entities

This frequency and ratios analysis added value to previously obtained insights by calling attention to certain relationships of which the company can take advantage, either for auditing purposes understanding if the behavior of the comprised entities is according to the expected, or to boost the relationship itself implementing methods to try to raise its market share.

3.4.7. Community Detection

The final part of the modeling of the providers' data was community detection. The goal was to find groups of providers highly connected/dense among themselves and divided by sparse spaces from the other groups. The ultimate objective was to constrain the analysis, focusing it on a certain group of nodes with characteristics that could, potentially, be important for the company. As observed when reviewing the existing literature on this subject, there are still some limitations on the community detection algorithms viable for bipartite graphs being the natural alternative using the two-mode projection of the network.

Consequently, the approach in this project included the implementation of algorithms, designed for unweighted unipartite structures, in the two-mode projection of the network, specifically on the physicians' side (in line with the business, it was determined that it had more value, exploring the possibility of engaging with them or applying upcoming insights in a new emerging company's program). Important to highlight that although the focus of community detection was on the physicians' side projection, the labels for the providers (*exec.*) side were still determined through the different algorithms to safeguard future developments where these may be needed, either for a combined approach or dual projection community detection. Nonetheless, in the end, it was also adapted one method that could be implemented on bipartite structures.

As presented in Network Statistical Measures, beforehand the local and global clustering coefficients of the network were measured, 0.143 for the bipartite graph and 0.804 for the unweighted projection G_1, indicating that the original network is sparse and the projection is dense, however the case of the latter value should be dealt carefully as it can be misleading (Latapy et al., 2008), as such these measures were used for reference reasons only.

The algorithms applied included the *Greedy Modularity Maximization,* the *Girvan-Newman,* the *Kernighan-Lin Bisection,* the *K-Clique,* the *Label Propagation (LPA),* the *Louvain Method,* and finally the *Bi-Louvain Method.*

Being this an unsupervised methodology, it was hard to define evaluation metrics, however for some algorithms was direct that the results were not according to the intended. Table 23 displays those cases and the criteria used for each case (specifically to reject some).

Table 23 – Community detection algorithms applied to the providers' data and criteria for rejection

Algorithm	Graph of application	Rejected?	Criteria for rejection
Greedy Modularity Maximization	(a)[2]	No	---
Girvan-Newman	(a)	Yes	Alone, one of the detected communities accounted for 96% of the physicians
Kernighan-Lin Bisection	(a)	Yes	This type of algorithm divides the network into two communities. In this case, one contained the most important elements regarding centralities and the other the least important – this algorithm was not analyzed further, as the conclusions would be identical to the ones obtained in the centralities' analysis
K-Clique	(a)	Yes	Limitation of not assigning to any community 21.74% (#145) of the referring physicians
Label Propagation (LPA)	(a)	Yes	Similar criterium to the *Girvan-Newman* algorithm
Asynchronous LPA	(a)	No	---
Louvain Method	(a)	No	---
Bi-Louvain Method	Unweighted Bipartite graph	No	---

After computing each algorithm and obtaining a community label for the network's nodes, the results were added to a dataset that contained several features that could be helpful in characterizing the different communities or restraining the analysis to a particular set of groups. Mainly, the provider's ID and name, the total number of distinct patients, the total number of claims associated with it (in

[2] (a): Unweighted unipartite projection of the original network onto the physicians' side

the case of providers *req.* it corresponds to the number of claims that contained a prescription made by the physician), the district, a Boolean variable indicating if the provider belonged to the set of "big groups", and the importance of each node given by the four centrality measures calculated (measured on the same network where the community detection was performed). Below the specifics of the algorithms that were kept are going to be detailed.

Starting with Modularity-based communities, specifically the Greedy Modularity Maximization algorithm, which assigns community labels based on the largest modularity (*Clauset-Newman-Moore greedy modularity*) values. This algorithm was applied using the package *community* from the *networkx* library. The parameters were set to their default values, in particular, a resolution equal to 1 (not favoring larger or smaller communities), a cutoff of 1 (meaning that even if the algorithm does not reach the modularity maximum it stops once it reaches this number of communities), and not defining a maximal number of communities that would keep the algorithm running even if modularity starts decreasing (*Greedy_modularity_communities — NetworkX 3.0 Documentation*, n.d.).

Through the Louvain community detecting heuristic method and based on modularity optimization, the network was divided into partitions following a two phases methodology. The resolution hyperparameter was set to 1, meaning that there was no favoring of larger or smaller communities, and the modularity gain threshold was defined as 0.0000001, that is, the algorithm would stop if the gain between two iterations was below this value.

Following with the Label Propagation algorithm with asynchronous label updating (through an application available in the previously mentioned package), several communities were obtained. The hyperparameters included were set to their default.

Now focusing on the extension of the Louvain algorithm for bipartite graphs, which was applied using a python's script created by Feng Liang (C. Zhou et al., 2018). This approach is a novel community detection method that has shown itself effective in identifying community structures in this type of network. The basis for the algorithm is its derivation of the gain of bipartite modularity and the agglomerative hierarchical clustering method that allows obtaining a complete partition structure.

After applying several algorithms, it was time to characterize the resulting communities and find common patterns within each partition. To facilitate the visualization of the several communities and let the business teams analyze them as they wished, it was created a *powerbi* dashboard where the user could choose the algorithm and the nodes' side to be examined. Additionally, several filtrations on the number of links, number of claims, number of elements in the communities, and on the physicians' entities were allowed (to focus the analysis on a more restricted number of communities). This methodology also served the purpose of evaluating the quality of the results, being a complement to the metrics that will be presented afterward. Various characteristics were used to describe the different groups, examples include the district, the centrality measures' values, the total billed value, the nodes' connections, and the claims number, among others. Table 24 displays a description of the several communities obtained by each algorithm (notice that although some communities obtained from different algorithms have the same name/label, these are not related). Also, Appendix F exemplifies some screens of the created dashboard. Besides, it is important to notice that other than understanding the elements of these groups, the purpose of community detection was also to gather insights into the structure of the network.

Table 24 - Main characteristics of the communities obtained by the several algorithms

Algorithm	Number of communities	Main Communities	Important communities (average degree-based)	Health units
Greedy Modularity Maximization	18 physicians' communities	2 main communities: A0: 55.47% (#370) – From *Lisboa* (70%), emphasis on the link between *Dr. RNAS* and *CCS* (in terms of the number of claims and invoiced value); A1: 40.93% (#273) – From *Porto* (38%); Remaining communities: 3.6% (#24)	A0, A1, A2, A3; the rest has similar importance	Only 6 communities have doctors linked to health units where they do not work
Asynchronous Label Propagation	19 physicians' communities	2 main communities: A0: 60.87% (#406) – From *Lisboa* (65.71%), emphasis on the link between *Dr. RNAS* and *CCS* (in the number of claims and invoiced value), also highlight for *Dra. MO*; A1: 34.48% (#230) – *Porto* (44.63%): being from *Porto* increases the probability of belonging to this group by 4 times; Remaining communities: 4.65% (#31)	A0, A1, A7, A12, A3; the rest has similar importance	Only 7 communities have doctors linked to health units where they do not work
Louvain Method	19 physicians' communities	4 main communities: A4: 45.43% (#303) – Lisboa (79.21%) – emphasis on the link between *Dr. RNAS* and *CCS* (in number of claims and invoiced value); A1: 36.73% (#245) – Porto (41.63%) – being from Porto increases the probability of belonging to this group by 3.8 times and being from Braga by 3 times; A2: 7.95% (#53); A8: 6.6% (#44); Remaining communities: 6.59% (#31)	A4, A2, A1, A8, A6; the rest has similar importance	Only 7 communities have doctors linked to health units where they do not work
Bi-Louvain Method	55 communities	4 main communities: Community 0: 17.11% (#192) [137 doctors + 55 health units]– Lisboa (81.77%) – emphasis on the link between *Dr. NC* and *CCDC* (in number of claims) and between *Dra. ARV* and *Lab Dr. JCAC* (in invoiced value); Community 3: 11.14% (#125) [76 doctors + 49 health units]– Porto (64%) – being from Porto increases the probability of belonging to this group by 9.9 times and being from Aveiro by 2.5 times; Community 25: 5.08% (#57) [30 doctors + 27 health units]; Community 11: 4.63% (#52); Remaining communities: 62.04% (#696)	Communities 5, 0, 12, 23, 15, and 40, among others	Only 6 communities have doctors linked to health units where they do not work

Resuming the evaluation of the algorithms, it is essential to highlight that this a blind procedure where there are no community labels of reference for the several nodes, so the evaluation needs to be internal, which became a challenge as the criteria that could be applied are more subjective and limited. First, as mentioned, the characterization of the communities was the first step of this evaluation – with it several algorithms were disregarded by not passing the quality test (Table 23). Furthermore, three other evaluation metrics were used and compared, mainly the average modularity, the number of communities, and execution time (Table 25).

Table 25 – Community detection algorithms evaluation

Algorithm	Number of communities	Modularity	Execution time
Greedy Modularity Maximization	18	0.331	0:00:04.735409
Asynchronous Label Propagation	19	0.307	0:00:00.010993
Louvain Method	19	0.339	0:00:01.653979
Bi-Louvain Method	55	0.803	0:00:05.858393

Analyzing the previous results, it is important to differentiate between the community detection on the projection (first three algorithms) and the one performed on the whole graph (last case). Besides, the bi-Louvain algorithm was computed for a larger volume of data by comprising the entire network.

Regarding the strength of the division of the network into partitions, the Louvain method is the one that seems more promising (although by residual differences). As for the Bi-Louvain method, the Bi-modularity value is quite high, indicating a strong community structure. In terms of execution, none of the methods presented constraints, but highlight for the Asynchronous Label Propagation which is efficiently performed. Nonetheless, if the universe under analysis becomes more comprehensive, the execution time will be vital. For the number of communities, the results are similar between the first three. Finally, by evaluating multilaterally the previous indicators, the *Asynchronous* LPA had better execution time but lower modularity, while the *Greedy Modularity Maximization* was the opposite. The *Louvain Method* seems a good compromise between the number of communities, modularity, and execution time. As for the *Bi-Louvain Method*, given the partitions' strength and number of communities obtained in under six seconds, it seems to be a good result in the context of these criteria. When it comes to community detection is a good practice to visualize the network according to the different communities, adding significance to the visual network analysis.

3.4.8. Network Visualization

Network visualization can also be valuable in an SNA project. On one side, materializes the object under study giving a sense of the structure of the society in question and how its different elements are related. On the other hand, although in its raw design, it is not easy to get relevant insights, if the right considerations are made (such as filtering links and elements, using coloring, and choosing the right layout) it can be an asset, which turned out to be checked in this project. Upon construction of the network, a preview of the network was already given (Figure 3.4 - Cardiology Network Visualization). At first, as observable, it is not possible to obtain direct insights, however, the fact that it was displayed with a layout given by the map of Portugal, already helped organize it.

So, to get a clearer view and gather additional knowledge, approach (A) (see Figure 3.12) consisted in:

1. Color each bipartition with a different color (the bipartite graph is a two-color graph), in this case, the set of providers *req.* was colored green and the providers *exec.* with blue.
2. Only connections having a number of shared claims above X were shown (from a grid of values, several tests were performed to understand what the most adequate value for X was, in the end, the chosen value was 20)

3. Also, to understand the magnitude of the node's importance, the size of the nodes was made proportional to their degree value, the more important the node, the greater the size
4. In terms of labelling, the names of the different providers were also displayed if their degree value was above a given threshold (in the upcoming visualizations the names will not be displayed for the sake of anonymity, but these were helpful)

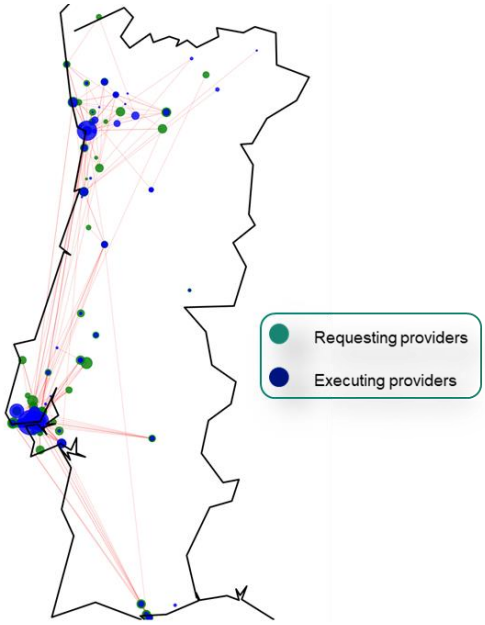

Figure 3.12 - Approach (A): Network Visualization

Approach (B) was to draw the network (see Appendix G) that contained the most important physicians (according to their degree centrality) and their respective connections (also combining points 1. and 3. mentioned above). Whereas approach (C) included displaying the network (see Appendix G) that contained the most important *MCDTs* providers (according to their degree centrality) and their respective connections with prescribing doctors (also combining points 1. and 3.).

From the previous illustrations of the network, from different perspectives, one valuable insight obtained was referred to the fact that many of the *MCDTs* are requested by doctors in regions further away from the place where they are performed (in another district) – e.g., many prescriptions requested in Lisbon are performed in other places like Algarve or Leiria. In First network insights, this possibility was raised, and these visualizations reinforce it, that is, there is a shortage of cardiologists that partner with the company within certain locations – a gap to be considered by the company in their providers' management process (for validation purposes the location of the patients involved in the different locations was analyzed, to know if, in fact, the *MCDTs* were performed at their place of origin).

Finally, approach (D) was an attempt to explore several different layouts, besides the map of Portugal. These included the following: *Fruchterman-Reingold* layout (force-directed algorithm), Circular layout (circular positioning), Random layout (random positioning), Shell layout (nodes positioned as concentric circles), Spring layout (same as *Fruchterman-Reingold* layout, but using edges as springs to keep nodes near), and Spectral layout (nodes positioned according to the graph Laplacian eigenvectors). Figure 3.13 exemplifies the raw application of these methodologies to graph G.

Although further explored, none of the previous layouts provided better visualizations than the previously presented approaches.

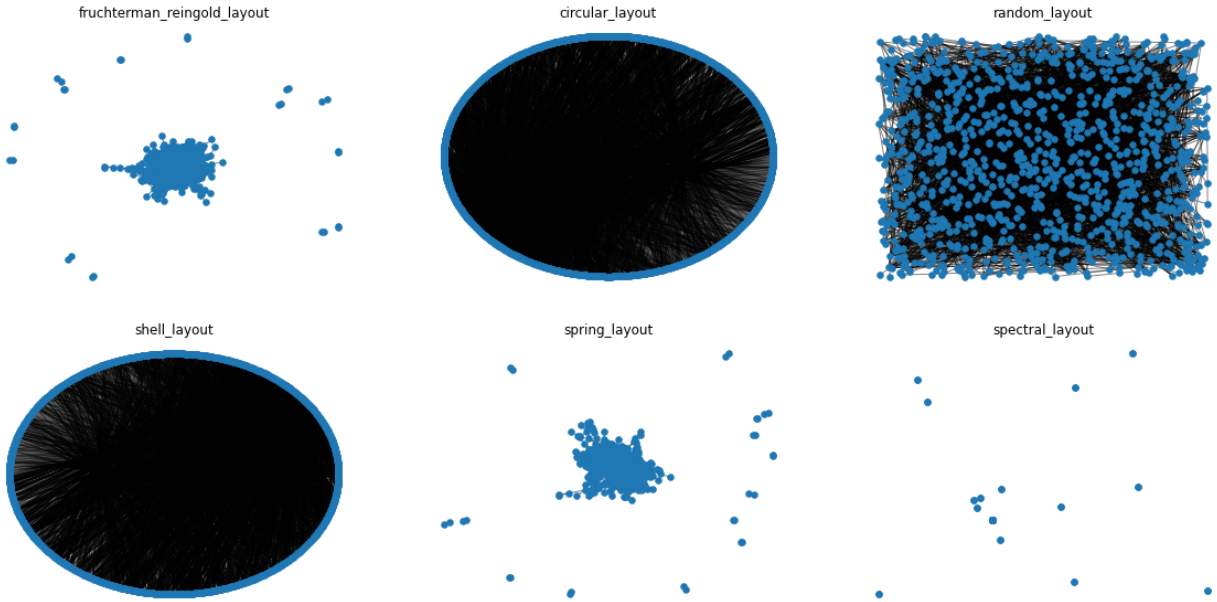

Figure 3.13 – Exploration of several network visualization layouts

Lastly, the VNA process of this project also comprised the visual analysis of the community detection results. Accordingly, Figure 3.14 depicts the network of physicians (two-mode projection) and its community structure obtained using the *Louvain* method. For simplicity, the network was displayed under the spring layout. From the picture, it is possible to determine what the communities more central are, as the nodes' size is proportional to their degree centrality. Additionally, this visualization can serve as a complementary evaluation criterium for the quality of the algorithm – it is possible to visualize a clear division among the main communities which may be an indicator of quality.

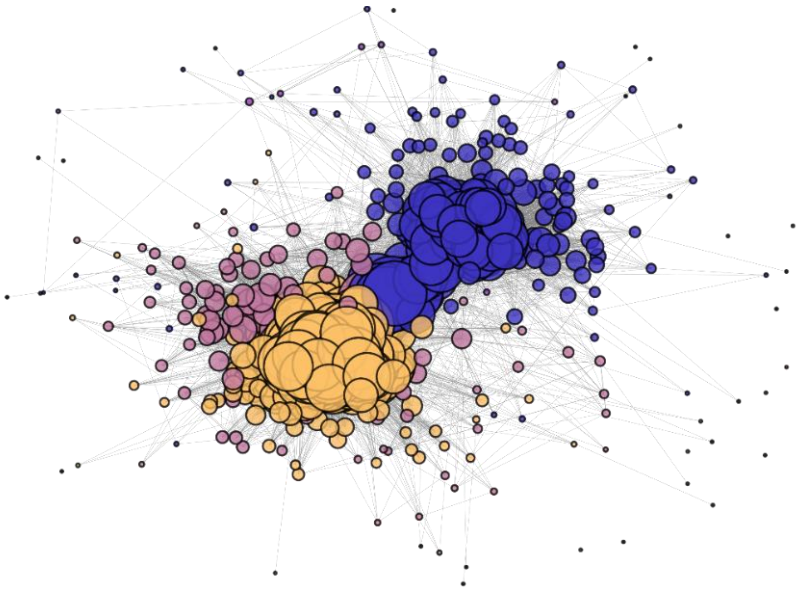

Figure 3.14 – *Louvain* method communities: Network visualization

4. RESULTS EVALUATION AND DISCUSSION

One fundamental part of a data analytics project is to evaluate both the results and the process that leads to them. As such, throughout this section, an overview of the modelling outcomes will be given along with a discussion of all the stages of this work.

Firstly, the business objectives for this work were to gather insights on the company's network of providers which was composed of connections between providers that request *MCDTs* cardiology-related and the providers that execute them, and attempt to understand what the main intervenients and relationships were, as well as if there were any niches or atypical cases that should be investigated or taken advantage of. After translating these points into a Data Mining language, SNA was performed.

From a combined approach between the analysis of a bipartite network that contained 3 175 distinct connections, and its dual projections, it was possible to turn the data into information. One example is the gap identified in the provision of specialized cardiology services - several individuals have consultations outside their hometown even though they can perform the prescribed *MCDT* there. The geographical location of the providers was an important aspect to consider, also giving a layout to draw the network on. The node-level statistical measures gave insights on the most important providers to keep inside the company's network of providers - from their management, the company can boost the business (for example by knowing that new providers (wherever their type is) that enter the network are more likely to join their counterpart if it is important). Although, the different centrality measures assigned slightly distinct rankings of importance according to their definition, when analyzing the distribution of one against the others the overall behavior showed that high values of one are usually associated with high values of the rest and vice-versa. Nonetheless, for the business context, the different definitions were presented along with the results, so that accurate choices can be made. One interesting insight is the fact that density is not necessarily linked to importance, some nodes locally were highly connected to their neighbors and in terms of importance were not on the top.

Examining the network as a whole, through network-level centrality measures and VNA, some perceptions of its structure were retrieved. One is the sparsity of the graph, which usually happens in real-world examples. Additionally, the degree distribution shows that this can be the case of a scale-free network, for having the majority of the nodes with a low degree, and a minority with elevated values.

To find tendential relationships, the two created metrics flagged various relationships between requesters and executers, however, most happened inside the same group, that is, there were cases where the physician worked at a place that also offered the possibility of executing the prescribed *MCDTs*, so naturally the patient would remain there. Also, most of these occurred inside the "big groups" of healthcare entities. Subsequently, an assessment outside these possibilities was made, from which potentially tendential relationships were signalized.

Although the clustering coefficient for the bipartite network was small, community detection was made to have a better understanding of the network properties but also to aim the analysis at a more restricted group of individuals. In this case, the focus was mostly on the physicians' side, for which three methods were applied to their unweighted projection (although to complement was also performed in relation to the projection on the other partition). As a supplement, the *Bi-Louvain* method was also employed for the whole graph. Looking into the results and taking the defined

evaluation criteria, the *Louvain* method seemed to provide the best combination of the number of communities, execution time, and average modularity, along with a good visual distribution of the main communities (although with close scores to the other algorithms). In terms of characterization of the several communities, all the algorithms led to similar conclusions - there is usually a community with a great number of physicians that practice in Lisbon, there is another that is from Porto, and so on, and these are usually connected to higher average degrees, which makes sense as the core of the business is within these locations. Even though the results do not seem the most insightful, they can help constrain future projects to certain niches of providers, or if there is a need to study a given attribute, its distribution across communities can be overviewed. Concerning the community detection on the bipartite network, this is still a field with great room to grow, however, the final average modularity indicates dense groups separated by sparse spaces – in the future a more detailed study can be made.

Overviewing all the steps from the data understanding until the modeling, naturally, some things could be improved, either because the available tools did not allow for their implementation, or because the time restrictions did not permit to go to such levels of detail by proceeding with an exhausting approach on every step of the project. However, some aspects that could be pointed out to be worked on in similar future projects include imposing restrictions on connections to be able to exclude spurious relationships between providers who treat the same patient but for opposite reasons (even if within the same specialty), this could be achieved using independent variables (to be determined) capable of controlling any spurious relationship that could appear. Also, instead of having, for example, the number of shared claims or patients or the number of occurrences as a link weight, alternative variables that assume this importance could be studied.

Additionally, when it came to the analysis of the centrality measures it was static over time. However, in future improvements it could be important to study how was the overtime evolution of the different providers, understanding if the positions that they occupied in terms of importance and relevance in the network shifted or not. This analysis could also be leveraged by dynamic graph visualizations, as opposed to traditional static graphs.

Furthermore, additional state-of-art methods could have been implemented, both to scrutinize the network at the node and network levels but also to approach community detection and perform a direct analysis of the bipartite structure in parameters other than those reviewed.

About the community detection, the main procedure was to apply several algorithms to the unweighted two-mode projection of the original graph, further analysis could be done to determine if there was any biased conclusion as a result. Also, some intellectuals argue that there could be some predilection in the clustering coefficient measured for this structure, as a dense network can be obtained because of certain groups of individuals having a high degree in the bipartite network, which does not necessarily mean the latter is dense as well (Latapy et al., 2008). In sum, the consequences of approaching certain aspects of this project using the two-mode projections can be further studied, for example, the implicit information loss or the possibility for biased results.

Making a global and final assessment of the current work, there is still room for improvement, but it was a learning journey, both at the academic level and at the company level, with results that had value for the business. Nonetheless, hopefully, it can serve as a basis for future projects and research.

5. CONCLUSION

The main objectives of this work included the study of the providers' network of a Portuguese health insurance company, along with understanding the feasibility of implementing SNA to approach it. Specifically, this network comprised relationships of providers that prescribed *MCDTs* to a given patient under the cardiology's specialty and the providers that end up executing them. Thus, this network was mirrored as a bipartite graph that encompassed two types of nodes whose relationships could only occur between actors from different vertices' partitions.

Key insights that the business aimed to obtain included understanding who the most important providers and relationships were while having a general overview of the topology of the network, their niches, and if there was any tendency in the connections or atypical findings. Ultimately, the business objective was to take the knowledge obtained to improve the firm's management of their in-network of providers. This undertaking consisted mainly of an exploratory analysis, which to answer to the previously enunciated purposes comprised a graph theory supported network statistical analysis at several levels of nodes' aggregation, the creation of metrics, network visualization, and the employment of community detection algorithms (see Modeling and results).

The process followed from the raw data sources to the final insights was long. In fact, before employing graph theory principles it was necessary to understand and transform the data - as common in companies of this sort, a vast knowledge of the business workflow is required to recognize the several assumptions needed, to know how to approach the several data sources, and to identify the right considerations at each step. Obviously, during it, the target outcome and the type of analysis to be made were always kept in mind. Also, throughout these stages, several insights were gathered, mainly related to the data characteristics, its elements, and geographical distribution, on top of all the others that were used in determining all the steps needed to prepare the dataset that would support the SNA.

After obtaining a final transformed dataset using *SAS EG*, this was fed into *python* where the network was constructed. The process of extracting information was achieved through a combined approach between the direct analysis of the bipartite graph and the dual projection of the network (one on each bipartition) with the particularity of performing the two-mode projection both into weighted unipartite networks and unweighted unipartite ones. The rationale behind this logic was to take advantage of several methods that were developed for different types of networks. Ultimately, by merging the various results, more robust conclusions were drawn. Through a statistical study, both at the network and node levels, it was possible to determine the most influential providers according to four different centrality perspectives: degree, closeness, eigenvector, and betweenness centralities, which although with different rankings followed similar patterns. Complementing it with the local clustering coefficient that evaluated the connectivity of each node to their neighbors, and other methods, a characterization of the individuals that belonged to this scale-free network was achieved. Also, the network structure was analyzed, to have an overview of the global behaviors and patterns across it.

When examining this type of data, one does not have to stick exclusively to SNA measures. Two additional metrics were developed to detect potentially biased relationships or determine important ones. Also, graph visualization had an important role in consolidating certain insights.

Finally, to find relevant groups of individuals, several community detection algorithms were applied with a focus on the physicians' side. Although atypical or surprising discoveries were not found, several attributes can be studied on levels more aggregated than the individual and more disaggregated than the whole graph. Nonetheless, the results can be classified as good if looking at the evaluation criteria.

The deployment of this project, in a business context, in addition to having involved the reporting and presentation of the work, will include the application of the results in different programs that are emerging inside the company, the creation of more detailed criteria for the management of providers within it, and the analysis of possible clinical protocols. As such, although for some people outside this scope, the results might be seen just as lists of providers and connections, they are much more than that. Naturally, there can be improvements, but this is a great stepping point for future employment. Thus, the contribution of this work can be both on the methodology and techniques utilized but also on the limitations (e.g., the projections' potential loss of information, the large data struggles, the networks' bipartivity constraints), and undone things.

This internship was a hard journey, but on my part, it generated a lot of learning, and value on the business side, and that is an achievement, even if some results might not be what was expected. Of some of the lessons learned, it should be noticed that when the objective is to innovate, and if studying a topic outside the normal scope, the time constraints should be considered carefully. Also, it is essential to have clear goals defined, and sometimes one looks for something atypical in the data, which is not. Likewise, SNA is a complex field with a lot of theories behind it.

Lastly, even outside the healthcare industry, SNA, despite its inherent challenges, can have a great impact on the study of societies from distinct environments.

6. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

In a project of this sort is hard not to point out some limitations or recommendations for future work.

First, the main limitation felt throughout the project, especially in the beginning, was the lack of computational power to process the network as a whole - even splitting the network by specialties, the execution times and memory consumption were too large. From the beginning this was an acknowledged risk, in fact, there was a contingency plan (which ended up being activated) consisting of focusing the analysis on the specialty of Cardiology. Another constraint on the tool level was the software availability – for example, the graph platform *Neo4J* is a powerful tool in SNA undertakings, however, the company did not have a license for its *Enterprise Edition*, so its community edition ended up being used not adding much value to the study as its resources were limited.

Another drawback was the timeline for the project. The lack of time meant that certain aspects of the data were not inspected as thoroughly as intended.

The restriction on cardiology can also be seen as a limitation as the full picture of the network is not observed.

Also, it was only possible to include providers inside the *Médis* Network (with agreements) due to the available data, and still, more features to characterize the providers could have been valuable.

Recommendations for future work involve the inclusion of external sources of data (only the company's data was used, leading to potentially missed information on providers' connections). Additionally, an SNA comprising an overview of the evolution of the network over time can be made (in this case the data was from one static year of activity) or even an extension of the scope of the project to other claims' years.

One interesting analysis to be made is the differentiation of networks according to the different customers' insurance plans, as the different contract conditions can affect their providers' choices.

Lastly, an improvement would be to have the full pathway of the customers' journey when evaluating the network – the outcome after the *MCDT* being performed (the return to the doctor, for instance) was not depicted.

7. REFERENCES

Aguilar, C. O. (2021). *An Introduction to Algebraic Graph Theory*.

Appel, A. P., Santana, V. F. de, Moyano, L. G., Ito, M., & Pinhanez, C. S. (2018). *A Social Network Analysis Framework for Modeling Health Insurance Claims Data* (Vols. 2022-March). IEEE Computer Society. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Aronson, B., Yang, K.-C., Odabas, M., Ahn, Y.-Y., & Perry, B. L. (2020). *Comparing Measures of Centrality in Bipartite Social Networks: A Study of Drug Seeking for Opioid Analgesics*. https://doi.org/http://dx.doi.org/10.31235/osf.io/hazvs

Arthur, R. (2019). *Modularity and Projection of Bipartite Networks*.

Azevedo, A., & Santos, M. F. (2008). *KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW* (H. Weghorn & A. P. Abraham, Eds.; Single). 978-972-8924-63-8.

Banerjee, S., Jenamani, M., & Pratihar, D. K. (2018). Algorithms for projecting a bipartite network. *2017 10th International Conference on Contemporary Computing, IC3 2017*, *2018-January*, 1–3. https://doi.org/10.1109/IC3.2017.8284345

Bang-Jensen, J., & Gutin, G. Z. (2009). *Digraphs*. Springer London. https://doi.org/10.1007/978-1-84800-998-1

Barrat, A., Barthélemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(11), 3747–3752. https://doi.org/10.1073/PNAS.0400087101

Bastian, M., Heymann, S., & Jacomy, M. (2009). *Gephi : An Open Source Software for Exploring and Manipulating Networks Visualization and Exploration of Large Graphs*. www.aaai.org

Bender, E. A., & Williamson, S. G. (2010). *Lists, Decisions and Graphs With an Introduction to Probability*.

Biggs, N. (1974). *Algebraic Graph Theory* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511608704

Bihari, A., & Pandia, M. K. (2015). Eigenvector centrality and its application in research professionals' relationship network. *2015 1st International Conference on Futuristic Trends in Computational Analysis and Knowledge Management, ABLAZE 2015*, 510–514. https://doi.org/10.1109/ABLAZE.2015.7154915

Bloch, F., Jackson, M. O., & Tebaldi, P. (2016). *Centrality Measures in Networks*. http://arxiv.org/abs/1608.05845

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10). https://doi.org/10.1088/1742-5468/2008/10/P10008

Borgatti, S. P., & Everett, M. G. (1997). Network analysis of 2-mode data. *Social Networks*, *19*(3), 243–269. https://doi.org/10.1016/S0378-8733(96)00301-2

Brinkmeier, M., & Schank, T. (2005). *11: Network Statistics*.

Burton, B., & Jesilow, P. (2011). How Healthcare Studies Use Claims Data. *The Open Health Services and Policy Journal*, *4*, 26–29.

Chandola, V., Sukumar, S. R., & Schryver, J. (2013). Knowledge discovery from massive healthcare claims data. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *Part F128815*, 1312–1320. https://doi.org/10.1145/2487575.2488205

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0 Step-by-step data mining guide*. DaimlerChrysler.

Chessa, A., Crimaldi, I., Riccaboni, M., & Trapin, L. (2014). Cluster analysis of weighted bipartite networks: A new copula-based approach. *PLoS ONE*, *9*(10). https://doi.org/10.1371/journal.pone.0109507

Clauset, A., Newman, M. E. J., & Moore, C. (2004). *Finding community structure in very large networks*.

Costa, C. J., & Aparicio, J. T. (2020). *POST-DS: A Methodology to Boost Data Science*.

Csikós, M., Hoske, D., & Ueckerdt, T. (2019). *Graph Theory lecture*.

da Fontoura Costa, L. (2022). *Discovering Patterns in Bipartite Networks*. https://doi.org/10.1101/2022.07.16.500294

Daugulis, P. (2012). A note on a generalization of eigenvector centrality for bipartite graphs and applications. *Networks*, *59*(2), 261–264. https://doi.org/10.1002/net.20442

Deo, N. (2016). *Graph theory with applications to engineering and computer science*. Dover Publications.

Dinh, L. T. N., Karmakar, G., & Kamruzzaman, J. (2020). A survey on context awareness in big data analytics for business applications. *Knowledge and Information Systems*, *62*(9), 3387–3415. https://doi.org/10.1007/S10115-020-01462-3

Du, D. (2018). *Social Network Analysis: Centrality Measures*. https://ddu.ext.unb.ca/6634/Lecture_notes/Lecture_4_centrality_measure.pdf

Eppstein, D. (2020). *Bipartite graph - Wikipedia*. Bipartite Graph. https://en.wikipedia.org/wiki/Bipartite_graph

Estrada, E. (2011). *The Structure of Complex Networks*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199591756.001.0001

Estrada, E., & Rodríguez-Velázquez, J. A. (2005). Spectral measures of bipartivity in complex networks. *Physical Review E*, *72*(4), 046105. https://doi.org/10.1103/PhysRevE.72.046105

Everett, M. G. (2016). Centrality and the dual-projection approach for two-mode social network data. *Methodological Innovations*, *9*. https://doi.org/10.1177/2059799116630662/ASSET/IMAGES/LARGE/10.1177_2059799116630662-FIG1.JPEG

Everett, M. G., & Borgatti, S. P. (2013). The dual-projection approach for two-mode networks. *Social Networks*, *35*(2), 204–210. https://doi.org/10.1016/j.socnet.2012.05.004

Fan, Y., Li, M., Zhang, P., Wu, J., & Di, Z. (2007). The effect of weight on community structure of networks. *Physica A: Statistical Mechanics and Its Applications*, *378*(2), 583–590. https://doi.org/10.1016/j.physa.2006.12.021

Fionda, V., & Pirrò, G. (2013). Querying graphs with preferences. *International Conference on Information and Knowledge Management, Proceedings*, 929–938. https://doi.org/10.1145/2505515.2505758

Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., & Lane, J. (2021). *Big data and social science : data science methods and tools for research and practice* (2nd ed., Vol. 63).

Freeman, L. (2000). Visualizing Social Networks. *Journal of Social Science*, *1*. https://www.cmu.edu/joss/content/articles/volume1/Freeman.html

Freeman, L. C. (2004). *The development of social network analysis*. 205. https://www.researchgate.net/publication/239228599_The_Development_of_Social_Network_Analysis

George, R., Shujaee, K., Kerwat, M., Felfli, Z., Gelenbe, D., & Ukuwu, K. (2020). A Comparative Evaluation of Community Detection Algorithms in Social Networks. *Third International Conference on Computing and Network Communications (CoCoNet'19)*.

Ghosh, S., Podder, M., & Sen, M. K. (2010). Adjacency matrices of probe interval graphs. *Discrete Applied Mathematics*, *158*, 2004–2013. https://doi.org/10.1016/j.dam.2010.08.018

Goldenberg, D., & Aviv, T. (2019). *Social Network Analysis: From Graph Theory to Applications with Python; Social Network Analysis: From Graph Theory to Applications with Python*. 1–2. https://doi.org/10.13140/RG.2.2.36809.77925/1

greedy_modularity_communities — NetworkX 3.0 documentation. (n.d.). Retrieved January 24, 2023, from https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.community.modularity_max.greedy_modularity_communities.html#networkx.algorithms.community.modularity_max.greedy_modularity_communities

Guichard, D. (2022). *An Introduction to Combinatorics and Graph Theory*. https://www.whitman.edu/mathematics/cgt_online/cgt.pdf

Guillaume, J.-L., & Latapy, M. (2004). *Bipartite Structure of all Complex Networks*.

Hart, M., Sherer, T., Blythe, M., Buck, A., Berdugo, M., Sharabi, K., Sparkman, M., & Sharkey, K. (2023). *O que é Power BI? - Power BI | Microsoft Learn*. https://learn.microsoft.com/pt-pt/power-bi/fundamentals/power-bi-overview

Hawe, P., Webster, C., & Shiell, A. (2004). A glossary of terms for navigating the field of social network analysis. In *Journal of Epidemiology and Community Health* (Vol. 58, Issue 12, pp. 971–975). https://doi.org/10.1136/jech.2003.014530

Haythornthwaite, C. (1996). Social network analysis: An approach and technique for the study of information exchange. In *Library & Information Science Research* (Vol. 18, Issue 4). https://doi.org/10.1016/S0740-8188(96)90003-1

Heaney, M. T. (2021). Tweeting #RamNavami: A Comparison of Approaches to Analyzing Bipartite Networks. *IIM Kozhikode Society & Management Review*, *10*(2), 127–135. https://doi.org/10.1177/22779752211018010

Heer, J., & Boyd, D. (2005). Vizster: visualizing online social networks. *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, 32–39. https://doi.org/10.1109/INFVIS.2005.1532126

Hoffman, M. (2021). 7 Network Visualization and Aesthetics | Methods for Network Analysis. In *Methods for Network Analysis*. https://bookdown.org/markhoff/social_network_analysis/network-visualization-and-aesthetics.html#layouts

Holme, P. (2019). *Firsts in network science – Petter Holme*. https://petterhol.me/2019/04/15/firsts-in-network-science/

KAGAN, J. (2022). *Health Insurance: Definition, How It Works*. https://www.investopedia.com/terms/h/healthinsurance.asp

Kernighan, B. W., & Lin, S. (1969). *An Efficient Heuristic Procedure for Partitioning Graphs*.

Klein, D. J. (2010). Centrality measure in graphs. *Journal of Mathematical Chemistry*, *47*(4), 1209–1223. https://doi.org/10.1007/s10910-009-9635-0

Konrad, R., Zhang, W., Bjarndóttir, M., & Proaño, R. (2020). *Key considerations when using health insurance claims data in advanced data analyses: an experience report*. https://doi.org/10.1080/20476965.2019.1581433

Kunegis, J. (2015). Exploiting the structure of bipartite graphs for algebraic and spectral graph theory applications. *Internet Mathematics*, *11*(3), 201–231. https://doi.org/10.1080/15427951.2014.958250

Labatut, V. (2015). Generalized Measures for the Evaluation of Community Detection Methods. *International Journal of Social Network Analysis and Mining (SNAM)*, *2*(1), 44–63. https://doi.org/10.1504/IJSNM.2015.069776ï

Landherr, A., Friedl, B., & Heidemann, J. (2010). *A Critical Review of Centrality Measures in Social Networks*. www.fim-rc.de

Landon, B. E., Onnela, J. P., Keating, N. L., Barnett, M. L., Paul, S., O'Malley, A. J., Keegan, T., & Christakis, N. A. (2013). Using Administrative Data to Identify Naturally Occurring Networks of Physicians. *Medical Care*, *51*(8), 715. https://doi.org/10.1097/MLR.0B013E3182977991

Larremore, D. B., Clauset, A., & Jacobs, A. Z. (2014). Efficiently inferring community structure in bipartite networks. *Physical Review E*, *90*(1), 012805. https://doi.org/10.1103/PhysRevE.90.012805

Latapy, M., Magnien, C., & Vecchio, N. del. (2008). Basic notions for the analysis of large two-mode networks. *Social Networks*, *30*(1), 31–48. https://doi.org/10.1016/j.socnet.2007.04.006

Lehman, E., Leighton, F. T., & Meyer, A. R. (2015). *Mathematics for Computer Science* . https://people.csail.mit.edu/meyer/mcs.pdf

Liebig, J., & Rao, A. (2014). *Identifying Influential Nodes in Bipartite Networks Using the Clustering Coefficient*. https://doi.org/10.1109/SITIS.2014.15

Linner, S., Lischewski, M., Richerzhagen, M., & Jülich, F. (2021). *PYTHON Introduction to the Basics*.

Liu, X., & Murata, T. (2009). How does label propagation algorithm work in bipartite networks? *Proceedings - 2009 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Workshops, WI-IAT Workshops 2009*, *3*, 5–8. https://doi.org/10.1109/WI-IAT.2009.217

Mandl, K. D., Olson, K. L., Mines, D., Liu, C., & Tian, F. (2014). Provider Collaboration: Cohesion, Constellations, and Shared Patients. *Journal of General Internal Medicine*, *29*(11), 1499. https://doi.org/10.1007/S11606-014-2964-0

Manning, C., Raghavan, P., & Schütze, H. (2009). *An Introduction to Information Retrieval*. Cambridge University Press.

Marcas do Grupo | Grupo Ageas Portugal. (n.d.). Retrieved January 15, 2023, from https://www.grupoageas.pt/marcas-do-grupo

Mascolo, C. (2011). *Social and Technological Network Analysis: Centrality Measures*. https://www.cl.cam.ac.uk/~cm542/teaching/2011/stna-pdfs/stna-lecture3.pdf

Medical Insurance Claim Procedure & How Does It Work? (n.d.). Retrieved December 25, 2022, from https://www.iffcotokio.co.in/health-insurance/what-is-medical-insurance-claim-procedure-and-how-does-it-work

Melamed, D. (2014). Community Structures in Bipartite Networks: A Dual-Projection Approach. *PLoS ONE*, *9*(5), 97823. https://doi.org/10.1371/JOURNAL.PONE.0097823

Meyers, L., Gamst, G., & Guarino, A. J. (2009). *Data Analysis Using SAS Enterprise Guide*. Cambridge University Press.

Murphy, P. (2019). *RPubs - Two-mode Networks in statnet*. https://rpubs.com/pjmurphy/542335

Naduvath, S. (2017). *Graph Theory-Fundamentals*. https://doi.org/10.13140/RG.2.2.16473.83040

Neal, Z. P., Domagalski, R., & Sagan, B. (2022). Analysis of Spatial Networks From Bipartite Projections Using the R Backbone Package. *Geographical Analysis*, *54*(3), 623–647. https://doi.org/10.1111/gean.12275

Neo4j, Inc. (2018). *Product Brief: The Neo4j Graph Platform*. www.neo4j.com

NetworkX — NetworkX documentation. (n.d.). Retrieved January 21, 2023, from https://networkx.org/

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(23), 8577–8582. https://doi.org/10.1073/PNAS.0601602103/ASSET/F86312F7-1DFE-4A0E-928F-6C1D4161BF34/ASSETS/GRAPHIC/ZPQ02306-2388-M07.JPEG

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

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, *69*(2), 026113. https://doi.org/10.1103/PhysRevE.69.026113

Oliveira, M., & Gama, J. (2012). An overview of social network analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*(2), 99–115. https://doi.org/10.1002/widm.1048

Ortiz-Arroyo, & Daniel. (2010). *Discovering Sets of Key Players in Social Networks*. 27–47. https://doi.org/10.1007/978-1-84882-229-0

Ostovari, M., Yu, D., & Steele-Morris, C. J. (2018). *Identifying Key Players in the Care Process of Patients with Diabetes Using Social Network Analysis and Administrative Data*.

Palla, G., Farkas, I. J., Pollner, P., Derényi, I., & Vicsek, T. (2007). Directed network modules. *New Journal of Physics*, *9*. https://doi.org/10.1088/1367-2630/9/6/186

Pavlopoulos, G. A., Kontou, P. I., Pavlopoulou, A., Bouyioukos, C., Markou, E., & Bagos, P. G. (2018). Bipartite graphs in systems biology and medicine: a survey of methods and applications. *GigaScience*, *7*(4), 1–31. https://doi.org/10.1093/GIGASCIENCE/GIY014

Pelechrinis, K. (2015). *TELCOM2125: Network Science and Analysis*.

Peng, Y., Zhang, B., & Chang, F. (2021). Overlapping community detection of bipartite networks based on a novel community density. *Future Internet*, *13*(4). https://doi.org/10.3390/fi13040089

Piccolo, S. A., Lehmann, S., & Maier, A. (2017). *Design process robustness: a bipartite network analysis reveals the central importance of people*. https://doi.org/10.1017/dsj.2017.32

Raghavan, U. N., Albert, R., & Kumara, S. (2007). *Near linear time algorithm to detect community structures in large-scale networks*. https://doi.org/10.1103/PhysRevE.76.036106

Referral - Glossary | HealthCare.gov. (n.d.). Retrieved December 25, 2022, from https://www.healthcare.gov/glossary/referral/

Robins, G., & Alexander, M. (2004). Small Worlds Among Interlocking Directors: Network Structure and Distance in Bipartite Graphs. *Computational and Mathematical Organization Theory*, *10*(1), 69–94. https://doi.org/10.1023/B:CMOT.0000032580.12184.C0

Rochat, Y. (2009). Closeness Centrality Extended To Unconnected Graphs: The Harmonic Centrality Index. In *ASNA*.

Rosen, K. H. (2019). *Discrete mathematics and its applications* (8th ed.). McGraw-Hill Education. https://imanulhuq.yolasite.com/resources/Discrete%20Mathematics%20and%20Its%20Applications%20-%208e%20%28Kenneth%20Rosen%29%20%5B9781259676512%5D_compressed-compressed.pdf

Saberi, M., Khosrowabadi, R., Khatibi, A., Misic, B., & Jafari, G. (2021). Topological impact of negative links on the stability of resting-state brain network. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-81767-7

Sahoo, R., Rani, T. S., & Bhavani, S. D. (2016). Differentiating Cancer from Normal Proteinprotein Interactions Through Network Analysis. *Emerging Trends in Applications and Infrastructures for Computational Biology, Bioinformatics, and Systems Biology: Systems and Applications*, 253–269. https://doi.org/10.1016/B978-0-12-804203-8.00017-1

SAS. (2017). *SAS Enterprise Guide: Fact Sheet*.

Sauter, S. K., Neuhofer, L. M., Endel, G., Klimek, P., & Duftschmid, G. (2014). Analyzing healthcare provider centric networks through secondary use of health claims data. *2014 IEEE-EMBS International Conference on Biomedical and Health Informatics, BHI 2014*, 522–525. https://doi.org/10.1109/BHI.2014.6864417

Saxena, A., & Iyengar, S. (2020). *Centrality Measures in Complex Networks: A Survey*.

Scellato, S., Mascolo, C., Musolesi, M., & Latora, V. (2010). Distance Matters: Geo-social Metrics for Online Social Networks. *Workshop on Online Social Networks*.

Scheinerman, E. R. (2012). *Mathematics: A Discrete Introduction* (R. Stratton, Ed.; Cengage Learning). 3rd.

Scott, J., Tallia, A., Crosson, J. C., Orzano, A. J., Stroebel, C., DiCicco-Bloom, B., O'Malley, D., Shaw, E., & Crabtree, B. (2005). Social Network Analysis as an Analytic Tool for Interaction Patterns in Primary Care Practices. *Ann Fam Med*, 443–448. https://doi.org/10.1370/afm.344

Serrat, O. (2017). Social Network Analysis. In *Knowledge Solutions* (pp. 39–43). Springer Singapore. https://doi.org/10.1007/978-981-10-0983-9_9

Sheshnath Jaiswal, A. (2022). Importance of Data Exploration in Data Analysis A Review Paper. *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT*, *2*(2). https://doi.org/10.48175/568

Snyder, J. (2019). Data Cleansing: An Omission from Data Analytics Coursework. *Information Systems Education Journal (ISEDJ)*, *6*, 17. https://isedj.org/;http://iscap.info

Sobre nós | Médis. (n.d.). Retrieved January 15, 2023, from https://www.medis.pt/sobre-nos/

Sun, H.-L., Ch'ng, E., Yong, X., Garibaldi, J. M., See, S., & Chen, D.-B. (2017). *A Fast Community Detection Method in Bipartite Networks by Distance Dynamics*.

Taheri, S. M., Mahyar, H., Firouzi, M., Ghalebi, E., Grosu, R., & Movaghar, A. (2017). HellRank: a Hellinger-based centrality measure for bipartite social networks. *Social Network Analysis and Mining*, *7*(1), 1–16. https://doi.org/10.1007/S13278-017-0440-7/FIGURES/7

Telatnik, M. (2020). *How To Get Started with Social Network Analysis | by Mitchell Telatnik | Towards Data Science*. https://towardsdatascience.com/how-to-get-started-with-social-network-analysis-6d527685d374

Top 13 Best Python Compiler For Python Developers [2023 Rankings]. (2023). https://www.softwaretestinghelp.com/python-compiler/

Tutorial Layouts Gephi. (2011).

Understanding Health Insurance. (n.d.). Retrieved December 25, 2022, from https://www.medicalbillingandcoding.org/health-insurance-guide/overview/

Venturini, T., Jacomy, M., & Jensen, P. (2021). *What do we see when we look at networks: Visual network analysis, relational ambiguity, and force-directed layouts*. https://doi.org/10.1177/20539517211018488

Vispute, P. (2020). Performance Evaluation of Community Detection Algorithms in Social Networks Analysis. *Bioscience Biotechnology Research Communications*, *13*(14), 388–393. https://doi.org/10.21786/bbrc/13.14/90

What is a provider? | healthinsurance.org. (n.d.). Retrieved December 25, 2022, from https://www.healthinsurance.org/glossary/provider/

What is the difference between a prescription and a referral for physical therapy? - ProAction Physical Therapy. (2021). https://proaction-pt.com/faq/what-is-the-difference-between-a-prescription-and-a-referral-for-physical-therapy/

Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a Standard Process Model for Data Mining*.

Yang, S., Keller, F. B., & Zheng, L. (2017). *Social Network Analysis: Methods and Examples*. SAGE Publications, Inc. https://doi.org/10.4135/9781071802847

Yannakakis, M. (1978). Node-and edge-deletion NP-complete problems. *Proceedings of the Tenth Annual ACM Symposium on Theory of Computing - STOC '78*, 253–264. https://doi.org/10.1145/800133.804355

Zaki, M. J., & Meira, W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*.

Zhang, P., Wang, J., Li, X., Li, M., Di, Z., & Fan, Y. (2008). Clustering coefficient and community structure of bipartite networks. *Physica A: Statistical Mechanics and Its Applications*, *387*(27), 6869–6875. https://doi.org/10.1016/J.PHYSA.2008.09.006

Zhou, C., Feng, L., & Zhao, Q. (2018). A novel community detection method in bipartite networks. *Physica A: Statistical Mechanics and Its Applications*, *492*, 1679–1693. https://doi.org/10.1016/j.physa.2017.11.089

Zhou, T., Ren, J., Medo, M., & Zhang, Y. C. (2007). Bipartite network projection and personal recommendation. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *76*(4). https://doi.org/10.1103/PhysRevE.76.046115

APPENDIX

APPENDIX A: CRISP-DM

The life cycle of a Data Mining project according to CRISP-DM and its dynamics between the six phases.

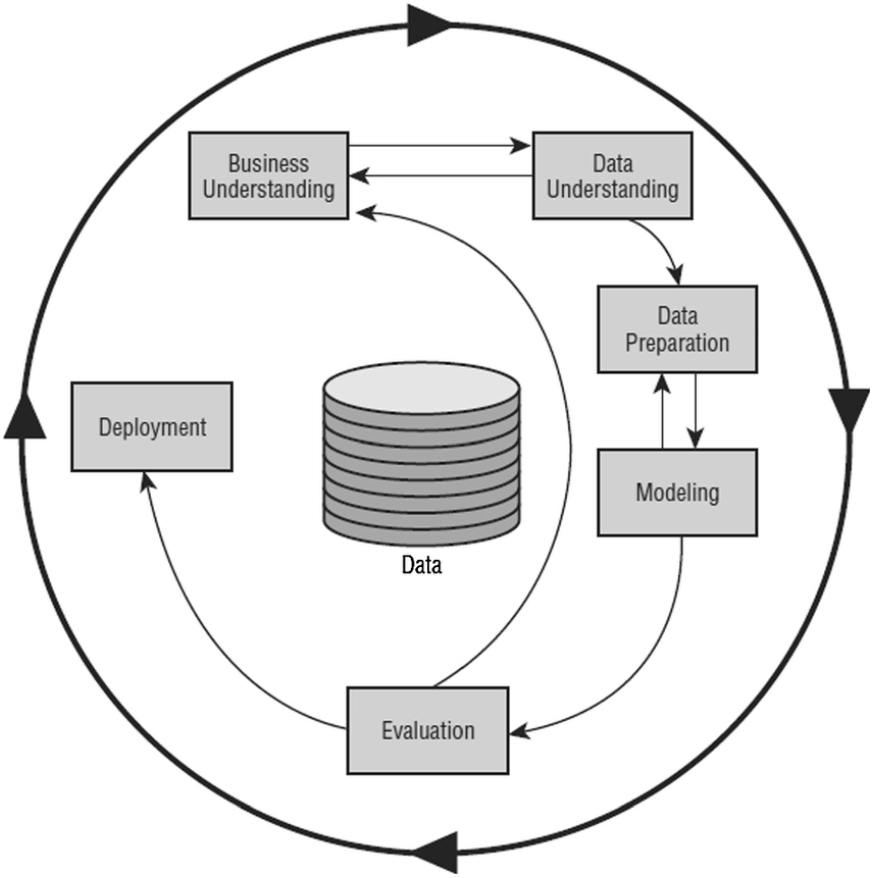

Figure 0.1 - Phases of the CRISP-DM [Source: (Dinh et al., 2020)]

APPENDIX B: METHODOLOGY

To better understand the process followed to answer the business objectives, Figure 0.2 mirrors the main steps comprised by this research.

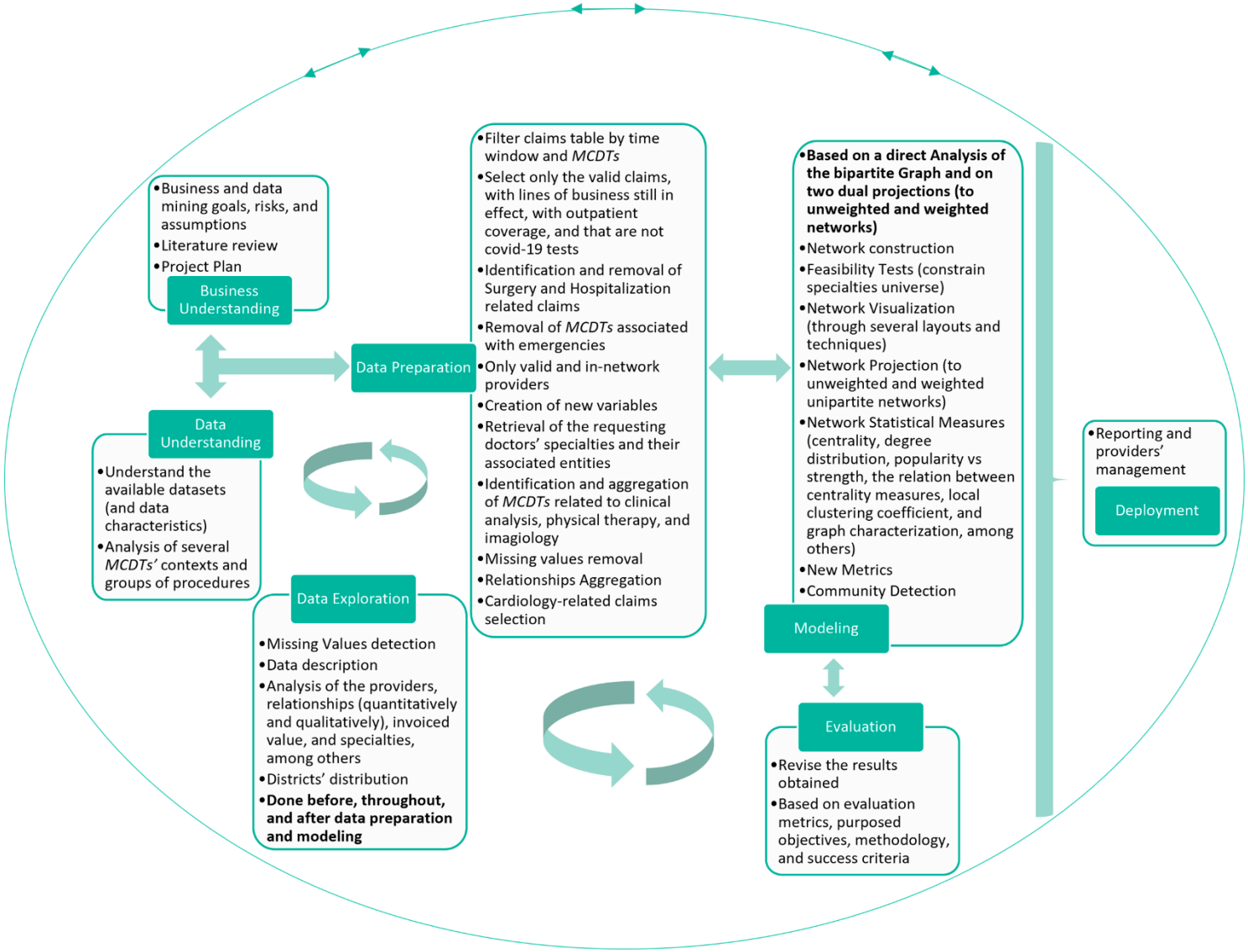

Figure 0.2 - Project's Methodology

APPENDIX C: PHYSICIAN'S SPECIALTIES

Before focusing the study on the Cardiology specialty, the claims data comprised 56 different specialties.

Table 26 - Physician's Specialties (top 7) distribution across the different claims

Specialty	Total no. of claims	% of the total no. of claims
Physical Medicine and Rehabilitation	266 619	37,9%
Gynecology-Obstetrics	91 659	13,0%
General clinic	66 214	9,4%
General Family Medicine	59 810	8,5%
Orthopedics	39 389	5,6%
Ophthalmology	22 165	3,1%
Cardiology	20 852	3,0%

APPENDIX D: UNIVERSE REDUCTION TESTING

To evaluate the type of universe reduction to be made, several tests on the specialties were made to study the need of focusing the analysis on a certain group or in a single specialty (see Table 27). Figure 0.3 reflects the complexity and number of relationships that the data encompassed, explaining the difficulties encountered before focusing on a single specialty.

Table 27 - Feasibility test results

Specialty	Running time	Feasible?
Whole data	*NA*	Unfeasible, given the technology available
Graph with the threshold defined at the 10% most relevant connections	*NA*	Unfeasible, even with the substantial data reduction
General Family Medicine	*NA*	Unfeasible, too many distinct connections
General clinic	*NA*	Unfeasible, too many distinct connections
Gynecology-Obstetrics	2h42min21s	Feasible. Although with a considerably high running time, it is estimated that further analysis could be done
Orthopedics	0h29min04s	Feasible
Internal medicine	0h24min36s	Feasible
Pediatrics	0h49min56s	Feasible
Cardiology	0h8min51s	Feasible

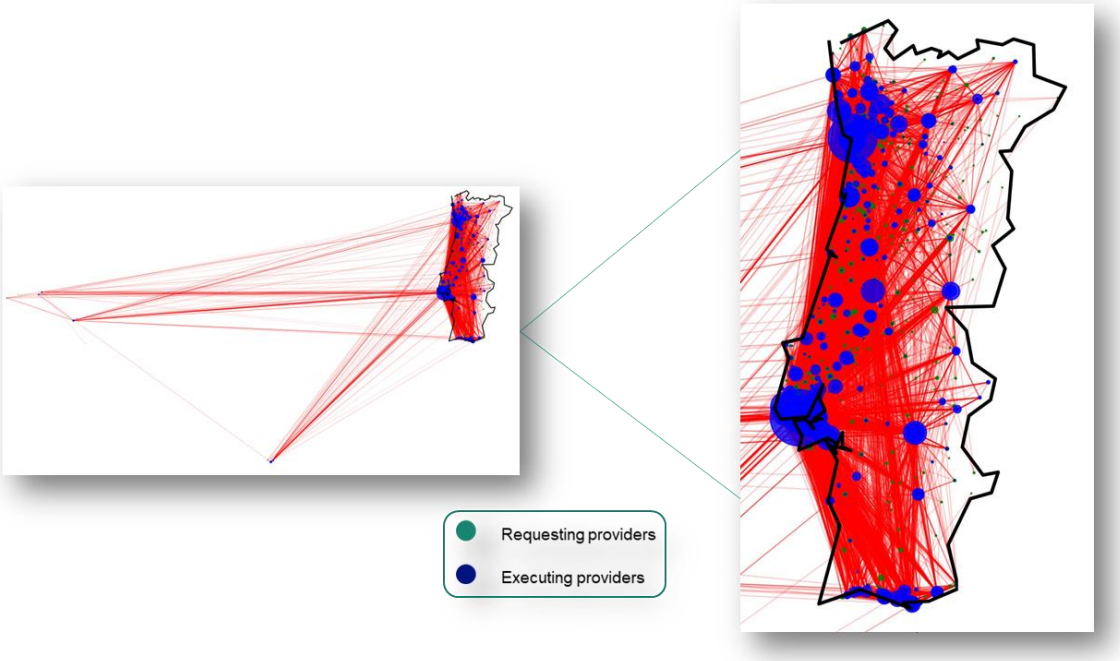

Figure 0.3 - Network Visualization (distributed across the map of Portugal), prior to the Cardiology filtration

APPENDIX E: CENTRALITY VALUES

To have an overview of the distribution of the centrality values across the several nodes of the network, the network was drawn using a centrality-based color map. These values were analyzed both in the one-mode projection of the physicians (see Figure 0.4) and in the network of executing providers (see Figure 0.5).

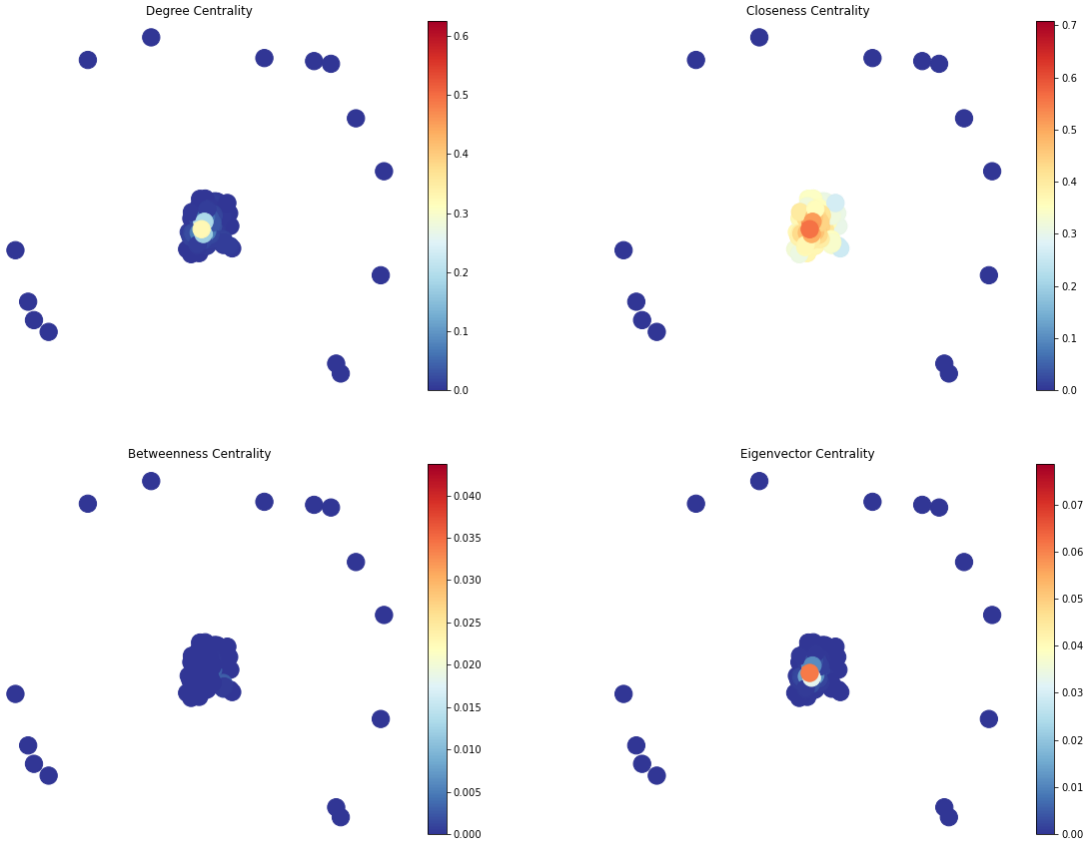

Figure 0.4 - Graph G_1: visualization using a centrality-based color-map

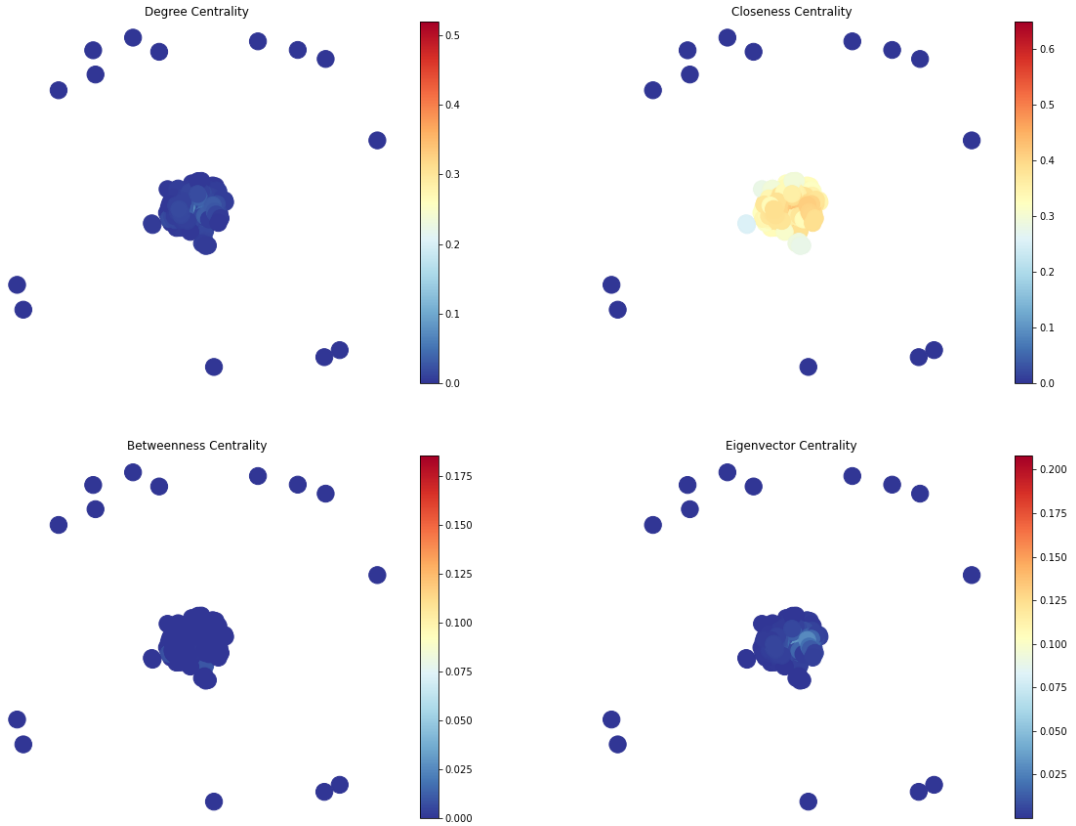

Figure 0.5 - Graph G_2: visualization using a centrality-based color-map

APPENDIX F: COMMUNITY DETECTION ALGORITHMS

To facilitate the visualization of the several communities and let the business teams analyze them as they wished, it was created a *powerbi* dashboard where the user could choose the algorithm and the nodes' side to be examined. Figure 0.6, Figure 0.7, and Figure 0.8 exemplify some screens of the created dashboard.

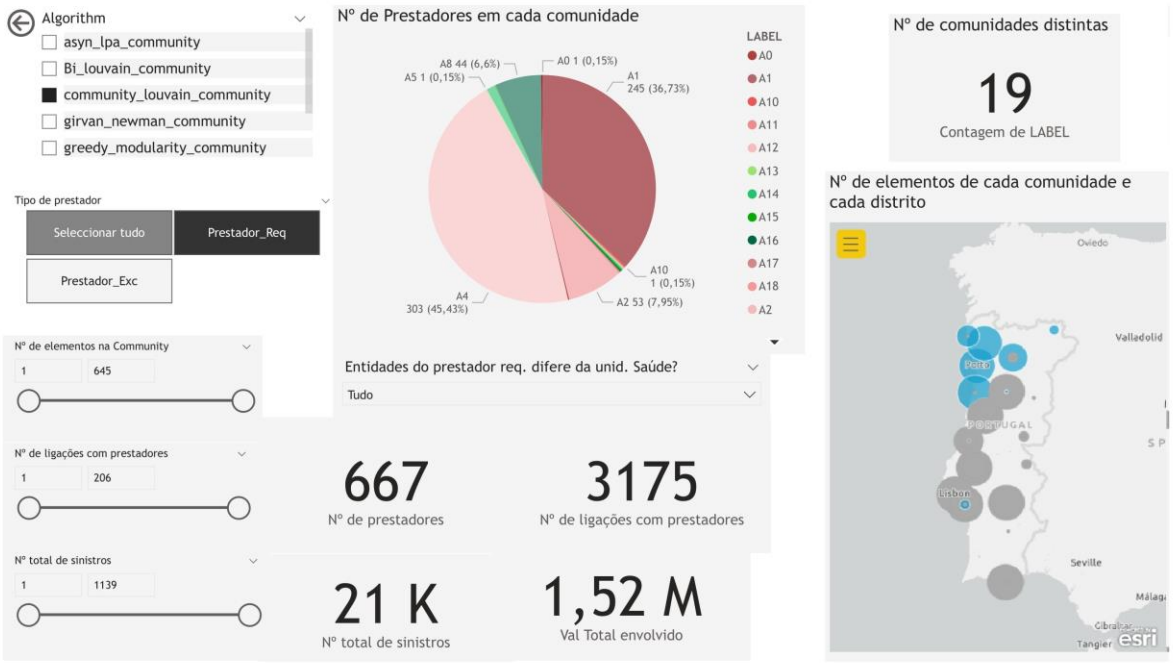

Figure 0.6 - Example 1 of the Dashboard developed for the analysis of the different community detection algorithms

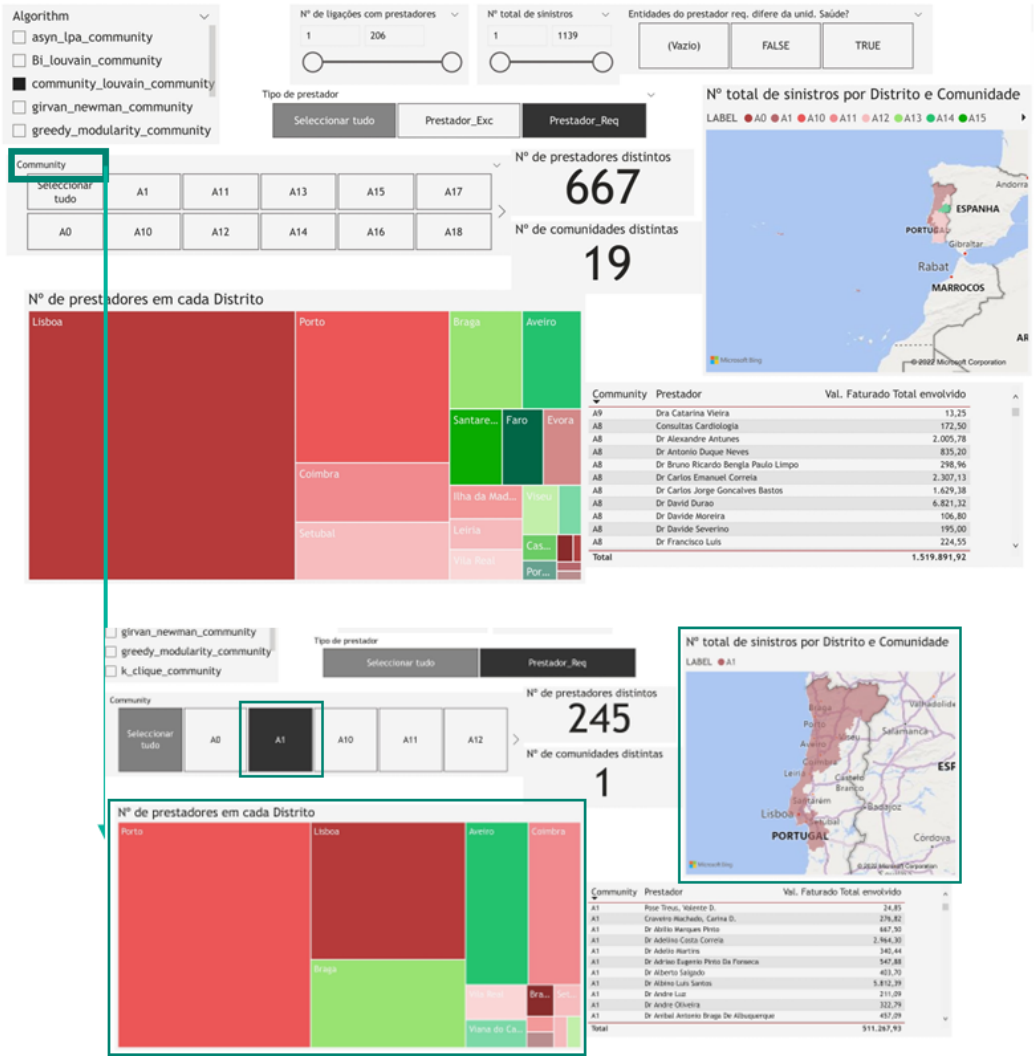

Figure 0.7 - Example 2 of the Dashboard developed for the analysis of the different community detection algorithms

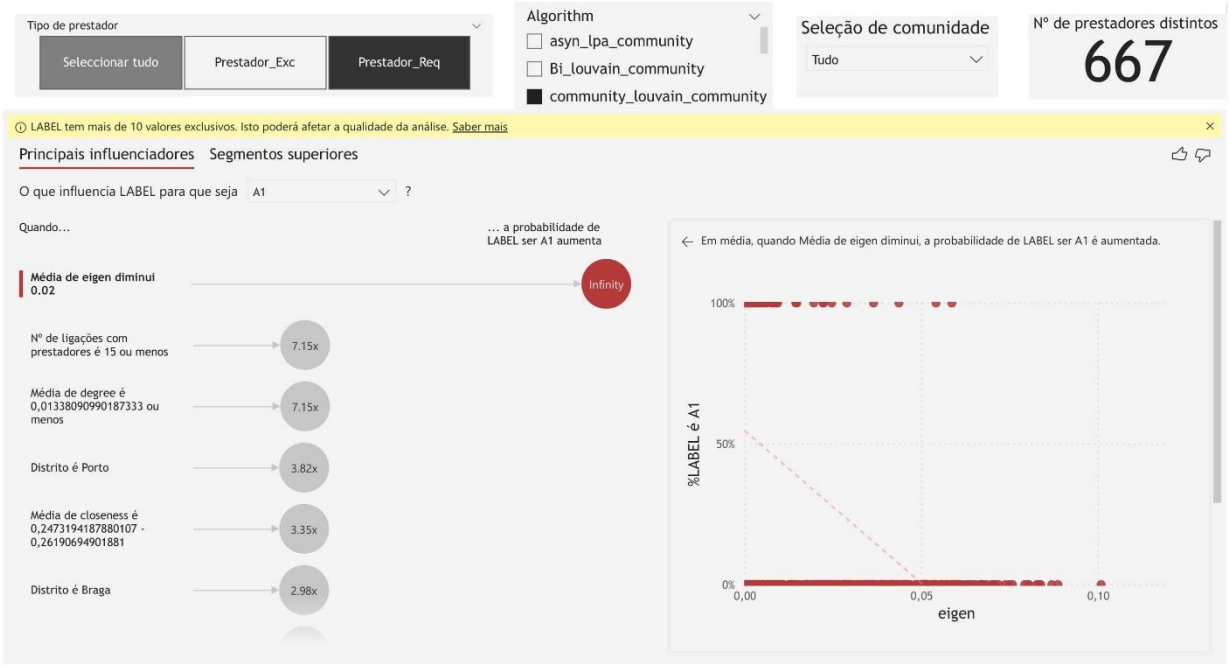

Figure 0.8 - Example 3 of the Dashboard developed for the analysis of the different community detection algorithms

APPENDIX G: NETWORK VISUALIZATION

To analyze and visualize the network of providers several approaches were taken. Figure 0.9 displays the most important physicians (according to their degree centrality) and their respective connections (the formatting assumptions are detailed in Network Visualization), whereas approach (C) included displaying the network that contained the most important *MCDTs* providers and their respective connections with prescribing doctors (see Figure 0.10).

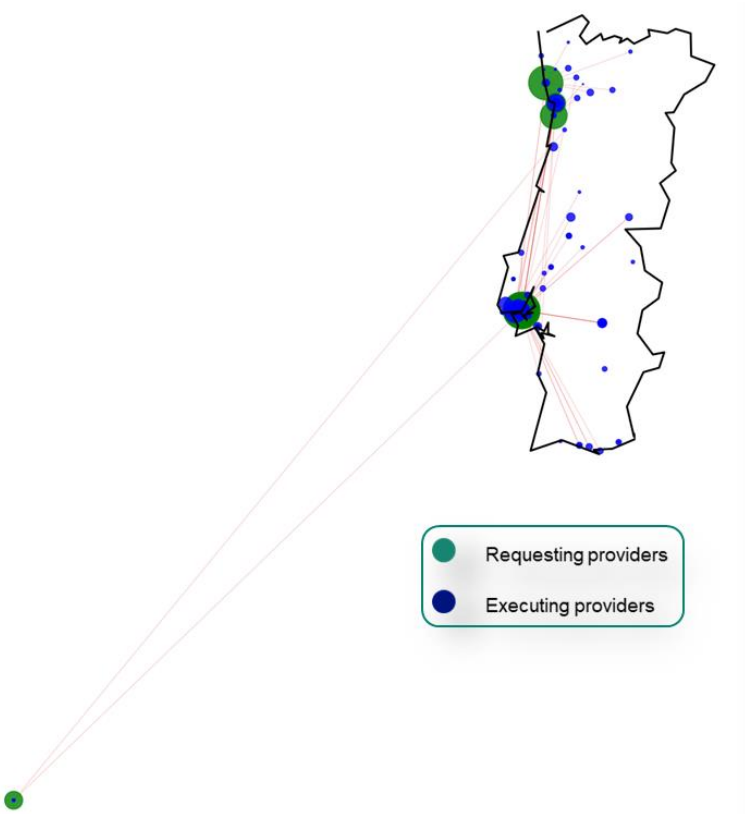

Figure 0.9 - Approach (B): Network containing the most important providers (*req.*) and their respective connections with healthcare units

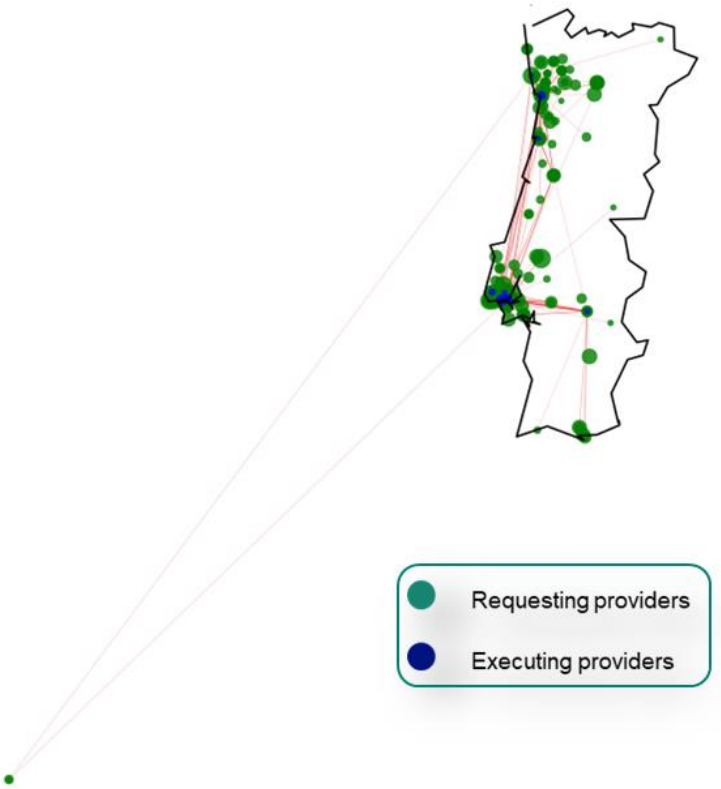

Figure 0.10 - Approach (C): Network containing the most important *MCDTs* providers (*exec.*) and their connections with prescribing doctors

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa