Ricardo José Monteiro Paiva

HOUSEHOLD IDENTIFICATION USING CALL RECORDS

1 2 9 0 UNIVERSIDADE D COIMBRA

UNIVERSIDADE D
COIMBRA

Ricardo José Monteiro Paiva

HOUSEHOLD IDENTIFICATION USING CALL RECORDS

**Dissertation in the context of the Master in Informatics Engineering,
Specialization in Intelligent Systems, advised by Professor Nuno Lourenço and
Professor Penousal Machado and presented to
Faculty of Sciences and Technology / Department of Informatics Engineering.**

October 2021

Faculty of Sciences and Technology

Department of Informatics Engineering

Household Identification Using Call Records

Ricardo José Monteiro Paiva

Dissertation in the context of the Master in Informatics Engineering,
Specialization in Intelligent Systems advised by Professor Nuno Lourenço and
Professor Penousal Machado and presented to
Faculty of Sciences and Technology / Department of Informatics Engineering.

October 2021

UNIVERSIDADE D
COIMBRA

This page is intentionally left blank.

Acknowledgements

First of all, I would like to thank my advisors, Professor Nuno Lourenço and Professor Penousal Machado, for their constant support and guidance throughout this last year. Thank you for this opportunity and for always being ready to help me whenever I asked for your advice. It is very comforting to be able to share ideas and openly discuss new ways of addressing a problem with such ease. You have helped me grow as a person, both personally and professionally.

Thank you to my parents and my brother for always caring for me and being there for me when I needed the most. It was with your unconditional support that I gained the strength to overcome the most difficult moments. I owe everything I have accomplished to you. Thank you for your patience and love.

Last but definitely not least, I would like to thank all my colleagues and close friends who have accompanied me over the years, and especially those who have filled these last years with so much joy, laughter and friendship. So, to Pedro, Duarte, João, Francisco, António, Jéssica, Beatriz, Rodrigo, Alexandre, Flávio and Guilherme. Thank you for accompanying me on this journey. Without you it wouldn't have been the same.

This work is funded by national funds through the FCT — Foundation for Science and Technology, I.P., within the scope of the project CISUC - UID/CEC/ 00326/2020 and by European Social Fund , through the Regional Operational Program Centro 2020, and the contract MEO-UC-ALB in the context of the protocol between Altice and the University of Coimbra.

This page is intentionally left blank.

Abstract

A good relationship with its customers is crucial to the success of any company. Nowadays, companies, especially those operating in the telecommunications industry, have increasingly invested in the relationship they have with their customers to ensure their satisfaction and, consequently, their retention. One of the techniques that have been used by companies in this area has been the creation of customer profiles through segmentation techniques, in order to understand their interests. Through the creation of these profiles, telecommunications companies are able to provide their customers with specialised services tailored to their needs and taste.

The main objective of this work is to evaluate the possibilities of the study of Call Detail Records with regard to customer segmentation by telecommunications companies. In concrete, we propose two models to analyse customers' social networks built on their communication records, with the aim of determining who are the most influential customers, as well as their social groups. In addition we present a way to classify customers according to their mobile operator.

These models allowed us to classify the most influential customers and the results of the study of their social groups confirmed the viability of detecting households by studying this type of data. Companies can use these results to adapt their strategies and marketing campaigns according to their customers' profiles to obtain a business advantage in such a competitive market.

Keywords

Call Detail Records, Social Network Analysis, Influencer Detection, Community Detection, Data Mining

This page is intentionally left blank.

Resumo

Uma boa relação com os seus clientes é fulcral para o sucesso de qualquer empresa. Nos dias de hoje, as empresas, principalmente aquelas que operam na área das telecomunicações, têm investido cada vez mais na relação que mantêm com os seus clientes de forma a garantir a sua satisfação e, como consequência, a sua retenção. Uma das técnicas que têm sido utilizadas nos últimos anos tem sido a criação de perfis de clientes através de técnicas de segmentação, de forma a perceber os seus interesses. Através da criação destes perfis, as empresas de telecomunicações podem fornecer aos seus clientes serviços especializados adaptados às suas necessidades e gostos.

O principal objectivo deste trabalho é avaliar as possibilidades do estudo de registos telefónicos no que diz respeito à segmentação de clientes por empresas de telecomunicações. Em concreto, propomos dois modelos para analisar as redes sociais de clientes construídas com base no seus registos de comunicação, com o objetivo de determinar quem são os clientes mais influentes, bem como os seus grupos sociais. Além disso, apresentamos uma forma de classificar os clientes de acordo com o seu operador móvel.

Estes modelos permitiram-nos classificar os clientes mais influentes e os resultados do estudo dos seus grupos sociais permitiram confirmar que é possível detectar agregados familiares através do estudo destas fontes de dados. As empresas podem utilizar estes resultados para adaptar as suas estratégias e campanhas de marketing de acordo com o perfil dos seus clientes de forma a obter vantagens comerciais num mercado tão competitivo.

Palavras-Chave

Registos Detalhados de Chamadas, Análise de Redes Sociais, Detecção de Influenciadores, Detecção de Comunidades, Exploração de Dados

This page is intentionally left blank.

Contents

This page is intentionally left blank.

Acronyms

CDR Call Detail Record. 2, 5, 7, 18–20, 22, 24

CRISP-DM Cross Industry Standard Process for Data Mining. 3, 5

CRM Customer Relationship Management. 23, 25, 26

ETL Extract, Transform and Load. 6

GDPR General Data Protection Regulation. 3, 20, 22

GPRS General Packet Radio Service. 24

HW Home and Work. 18

IMP Influence Maximization Problem. 17

NIC Client Identification Number. 25

PPP Post Paid Plan. 24, 31

SNA Social Network Analysis. 3, 7, 17, 19

This page is intentionally left blank.

List of Figures

This page is intentionally left blank.

List of Tables

This page is intentionally left blank.

Chapter 1

Introduction

The telecommunications industry today is characterized by a persistent search for loyal customers. However, this is a challenging task because of the increased competition based essentially on daily introduction of new products and services, on increasingly aggressive pricing strategies and by the higher level of customer demand, which is increasingly informed and has higher expectations. Thus, companies operating in this area have to search for solutions that enable them to obtain competitive advantages to leverage their business and competitiveness.

One of the ways in which companies have invested their resources in recent years is in how they manage customer relationships. Building a good relationship with customers leads to their satisfaction with the services provided, contributing to their retention and eventually leading to business growth. With this objective in mind, companies have to know their customers and become aware of their preferences and demands to make customer centered decisions and develop superior services and products that meet the identified needs.

By ensuring the satisfaction of current customers, organisations can unintentionally generate a wave of marketing buzz which is a type of marketing where satisfied customers not only buy more, but also make positive recommendations to other people [2]. A satisfied customer can attract many other customers just by communicating with each other. However, negative recommendations can have a much greater impact and drive away potential ones. It is therefore necessary to segment customers in the right way to assess the best approach for each type of customer.

Generally, the knowledge that companies have about their clients is done by collecting data from a range of different communication channels including the company's website, smartphones, email, marketing materials and social media. Through the analysis of usage patterns, customers are organized into groups based on common characteristics. These groups, also called segments, arrange customers who tend to respond similarly to different marketing strategies. With this type of information, companies can adapt the way they establish communication, how they communicate and what types of products are more adequate for those groups of clients.

Smartphones are becoming an essential object of our daily routine, especially with the growing speed of technological advancement. Through attractive and powerful applications, they can do almost everything and their availability and usage have been increasing. They also allow the recording of our activity in terms of use and location. As a consequence, large amounts of valuable data are generated every day and collected by different companies allowing the study individuals' usage patterns.

For each voice call, telecommunication companies store records that contain information about the success of the call, duration, numbers involved, among others. Most commonly known as Call Detail Records (CDRs) [3], this type of data was, initially, only collected for billing purposes and traffic management but, when companies found that this data could be applied in other areas, they started studying their clients through their calling patterns and calling locations.

In this work we explore the possibilities of studying Call Detail Records with regard to customer segmentation. Specifically, we intend to explore whether it is possible to detect relationship groups and measure the level of customer influence based on these sources.

This work is part of a project established between the University of Coimbra and a telecommunications company with the aim of developing new and innovative tools to support and improve the company service.

1.1 Motivations

The need to know more about the client and the way he relates to others are the two main motivations behind this work.

Firstly, if through the study of CDRs we can segment customers according to their relationship groups, whether these are social or family groups, it allows telecommunications companies to analyse the elements that constitute them and better identify the most appropriate products to offer. By offering this type of customised products, companies avoid user fatigue, which contributes to a better relationship with the customer.

In addition, if we can infer the most influential customers, we can understand how people relate to each other and how the flow of information occurs. This makes the job of marketers much easier. By creating targeted campaigns, these individuals can be used as a means of reaching a large number of people in order to advertise a particular product.

Our objectives are naturally connected to these motivations. The primary objective of this work is to answer the question: “Is it possible to identify relationship groups through CDRs analysis?”. The identification of these groups emerges in the literature through the analysis of networks generated on the most famous social platforms available such as Twitter and Facebook [4, 5, 6]. In this work, through the use of CDRs, we study the application of social network analysis techniques for the detection of relationship groups and analyse how the detected groups can be considered households. To do this, we build networks based on the telephone records making it possible to study the relationships between individuals.

In addition to this, these networks also allow individuals to be studied individually. Therefore, the question arises: "How can we identify the most influential elements in a network of interactions?". Influence can be measured in different ways. A common reader may assume that if a certain customer receives a lot of calls he or she can be considered as an individual with a lot of influence. This hypothesis may not be entirely correct, and therefore, we propose a way to measure the level of influence of the different individuals that make up a network based on different social network analysis metrics.

1.2 Contributions

This work resulted in several contributions, which are summarized below.

1. Development of a model that calculates the level of influence of individuals belonging to a network of interactions. The proposed model consists of a set of centrality metrics that evaluate different characteristics of the network and assign an influence score to each individual belonging to that network.
2. Recursive Louvain Algorithm, which corresponds to an adaptation of the state of the art Louvain algorithm. With this model we were able to divide the interaction networks created into relationship groups. These groups were then evaluated using a dataset containing the relative addresses of a set of customers. The results showed that there is a set of customers who share similar relative addresses and were classified as being in the same relationship group. We consider that customers who share these results, can be considered as members of the same household.
3. Classification of customers according to their mobile operator. Through the call detail records we are able to obtain information on which operator a certain client is associated with.

1.3 Document Structure

In the first chapter we introduced the problem addressed by this work and presented the main contributions. The remainder of this document is organised as follows.

Chapter 2 introduces an overview of the Cross Industry Standard Process for Data Mining (CRISP-DM) describing all its constituent phases. In addition, this chapter presents key Social Network Analysis (SNA) concepts and algorithms, focusing on centrality metrics and the most commonly used community detection algorithms. We also review work related to social network analysis, as well as the use of call detail records. At the end we briefly address the General Data Protection Regulation (GDPR) topic.

Chapter 3 presents the research objectives and details the datasets used and the transformations that had to be made. In addition, it describes the exploratory data analysis performed on the provided dataset.

Chapter 4 describes the influencer detection model. This chapter outlines its architecture and how the results are stored and analysed, describing the visual tool developed. In addition, it is also presented how the classification of customers based on mobile operators is done.

Chapter 5 of this document provides a description of the community detection model. This chapter discusses the decisions that had to be made and analyses the results obtained. The validation process developed with the purpose of identifying elements belonging to the same household is also presented.

Chapter 6 gathers the main conclusions and lessons learned, and also looks into possibilities for future work that come from this work.

This page is intentionally left blank.

Chapter 2

Background

Research on Call Detail Records (CDRs) has been increasing in the last few years, with different researchers applying different techniques in order to obtain information about individuals. Relationships and moving patterns are two examples of data that can be extracted from this type of records. This data can then be applied in different areas in order to obtain commercial advantages and improve people's life quality.

This chapter provides an overview of some important concepts and techniques for this work. We begin by describing the Cross Industry Standard Process for Data Mining by taking a brief look at all its constituent phases. Then, we make a short description of the type of information that Call Detail Records provide. Furthermore, some key concepts regarding social network analysis are also presented, namely the analysis of some network metrics and some community detection algorithms. A review of previous work in this topic is also included, with a special focus on the detection of important locations and the study of relationships between individuals. In the end, the chapter also includes a short analysis of the General Data Protection Regulation and some techniques that allow personal data to be used without violating the rules imposed.

2.1 Cross Industry Standard Process for Data Mining

Cross Industry Standard Process for Data Mining (CRISP-DM) [7] is an open standard process model that describes common approaches used by data mining experts to solve data mining problems and turn large amounts of data into knowledge to support decision making. With this methodology any professional with more or less experience can have a complete plan for carrying out a data mining project. CRISP-DM breaks down the life cycle of a project into six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment.

Due to its cyclical nature, this model can be run in a non-strict manner and it is possible to move back and forth between different phases if any questions arise.

The first phase is called **Business Understanding** and it may be the most important one [8]. It focuses on understanding the goals of the project from a business perspective, converting this knowledge into a data mining problem definition. It is also in this phase that a preliminary plan to achieve the objectives is produced, and it is necessary to define various aspects such as requirements, risks and possible solutions, assumptions, success criteria and evaluate the available data in order to ensure that the project produces a

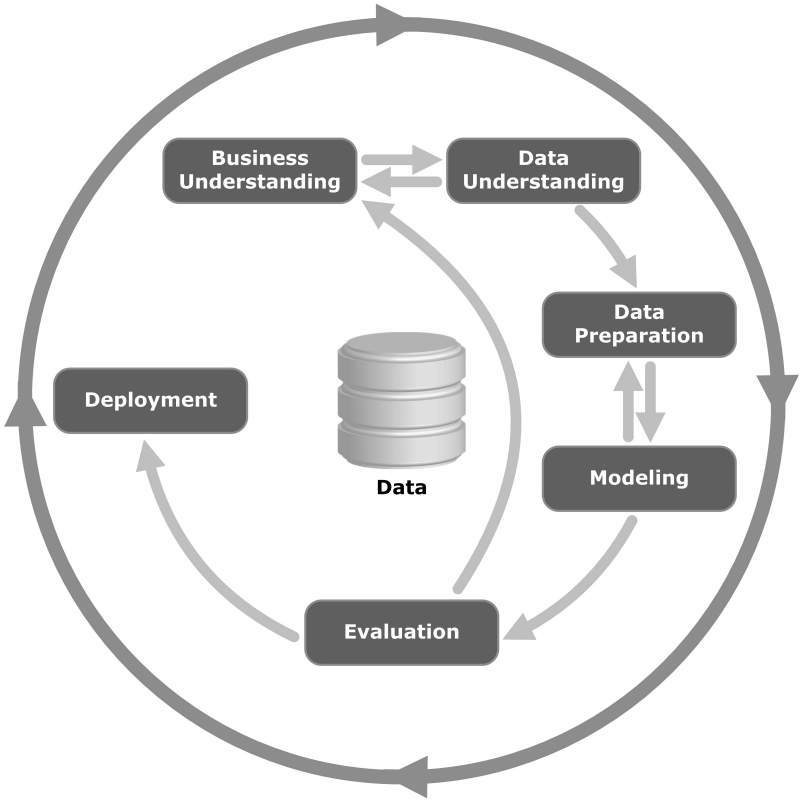

Figure 2.1: CRISP-DM Process Diagram

result according to what is intended.

The second phase, as the name suggests, is related to familiarization with the data. The **Data Understanding** phase starts with an initial data collection and analysis so that its quality can be assessed. The quality of the data can be evaluated based on several factors such as data completeness, relevance and value distribution. At the end of this phase there is already knowledge about the available data and all the transformations that will need to be performed to extract the best value and make it ready for use.

The **Data Preparation** phase covers all activities to construct the final dataset. In other words, it involves an Extract, Transform and Load (ETL) process that turns the data into something useful for the model to be developed. Included in this phase are tasks such as table, record and attribute selection, as well as data transformation and cleansing. This step can be time consuming and can get complex when working with large amounts of data.

Modeling is the core of any machine learning project. There are usually different techniques that can be applied to the same problem, and it is at this stage that these techniques are selected and applied. If necessary, the development team can go back to the data preparation stage to make some adjustments to bring the data in line with the chosen technique.

The **Evaluation** phase complements the modeling phase since it is necessary to evaluate the results of each of the models in order to select those that produce the best results. Taking into account what was defined in step one, if the results are not as expected, the methodology allows the development team to go back to the first step to understand why the results are not in accordance with what was defined.

The last phase corresponds to the **Deployment**. Usually the knowledge acquired by the model is organized and presented in a way that the customer can use. When this phase is reached, the project should have achieved its objectives. Depending on what was initially defined, the product of this phase can be as simple as generating a report or as complex as implementing the whole model on the client side.

2.2 Call Detail Records

CDRs are routinely collected by telecommunication providers in order to detect congested cell towers in need of additional bandwidth [3] and to bill customers for cellular usage. As such, they are a powerful source of information.

For each communication between individuals, the mobile operator keeps a CDR containing the type of call, phone numbers involved in the transaction, time, duration, cell tower ID, among others. The figure 2.2 shows an example of this type of data.

day_code	hour_code	min_sec	card_code	caller_num	called_ num	cell_code	call_dur	calltype	origdest	country_ code
20200901	0	11	19362	469917	155522	30421	87	0	0	351
20200901	0	127	28638	60063	376542	30375	49	0	3001	351
20200901	0	135	772	375140	2924	30371	34	3	5559	-1
20200901	0	245	30475	426956	11464	39454	58	0	5562	351
20200901	0	343	30434	17181	224475	34790	57	2	3001	351

Figure 2.2: Call Detail Records Example

The content of the communication is not revealed through the CDRs since they simply store properties related to the communication. They can also include SMS records or any other official communication transaction. For example, CDRs can be combined with the physical addresses of the cell towers to obtain the approximate location of where a particular call was made or received.

This type of data can be used to create networks, which could be analysed using network analysis techniques. Through this analysis and considering the amount of data that is generated every day, telecommunication companies and other entities can investigate the relationship level between persons, their mobility patterns, population distribution in different locations, among others.

2.3 Social Network Analysis

Social Network Analysis (SNA), also known as Network Science, is a knowledge area that uses networks and graph theory to understand relationships, interactions and communication patterns between elements represented as nodes and the connections between elements as links. These networks are based on relational data and can be applied to various scientific fields [9].

Representation

Graph theory is a mathematical field that studies graphical structures called graphs. They are composed of two fundamental components: nodes and edges. A node is an intersection

point which can represent a wide variety of individual entities (e.g., a person, a location, an organization). On the other hand, an edge is a link between two nodes and it can represent several kinds of relationships between entities. Formally, a graph G consists of a non-empty set V of vertices and a set E of edges, being defined as G=(V,E). Figure 2.3 shows an example of this structure.

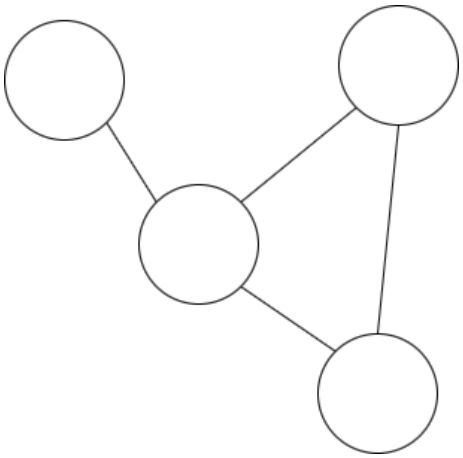

Figure 2.3: A Simple Graph

There are two types of edges: directed and undirected. Directed edges are applied from one node to another with a starting node and an ending node. Undirected edges are the opposite of directed edges because there is no clear starting and ending node (Fig. 2.4). Graphs can be classified according to the direction of their links. Undirected graphs are graphs whose edges do not have a defined direction and connect unordered pairs of vertices. Contrarily, directed graphs (also called Digraphs) use directed edges making the order of the vertices they connect matter.

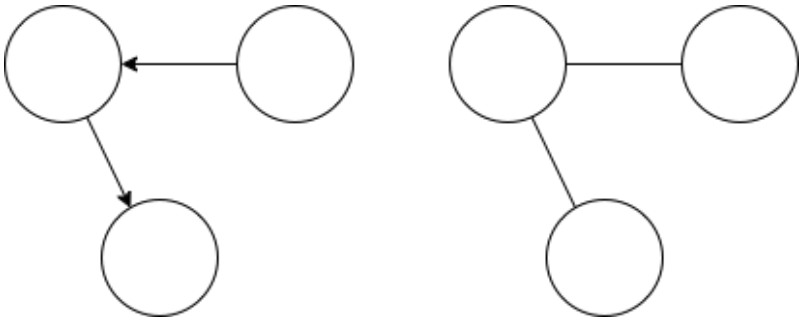

Figure 2.4: Directed and Undirected Edges

In addition to this, edges can have an extra feature which is the weight. An edge weight is the number of times that edge appears between two specific nodes. Generally, in social networks, the weight of a link could represent the duration, emotional intensity or frequency of interaction between two nodes [9].

To better understand the next section, it is important to define the adjacency matrix, which is a structure that is widely used in graph representation. The elements of this matrix indicate whether a pair of nodes is adjacent in the graph. Besides, this structure can also represent the weight information. In undirected and unweighted graphs, the adjacency matrix is binary and symmetric. For a pair of nodes (i, j), the entry a_{ij} of the matrix is 1 when exists a edge between the two nodes. Otherwise, it is 0. On the other

hand, for directed and weighted graphs, the entry a_{ij} of the matrix store the weight value of the edge between i and j. In this case, the adjacency matrix is non-symmetric. Figure 2.5 [9], illustrates an example of the adjacency matrix of an directed and unweighted graph.

Vertex	A	B	C	D
A	-	0	0	1
B	1	-	0	1
C	0	0	-	0
D	0	1	1	-

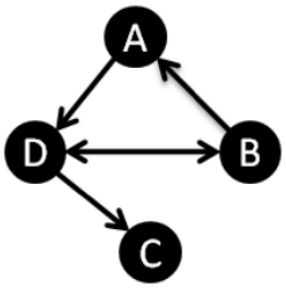

Figure 2.5: Adjacency Matrix Example

Metrics

Different metrics were developed in order to provide insights about the structure of the network and understand how the different actors are linked to each other. According to this, they can be divided into two groups: centrality measures and network level measures. The former attempts to examine the level of a given node in relation to the network using centrality procedures. The latter provides a more general level of information, making it possible to identify hubs or communities inside the network.

Centrality Measures

As stated above, centrality measures determine the importance of a node in relation to the network. A node with a high centrality value represents an actor with a high social strength meaning that is well connected to a large proportion of the remaining network nodes. Throughout the literature, this type of actors are also known as network influencers [1].

The **degree** of a node k_v is the number of edges the node has or, in a similar way, the number of neighbours. It is important to define a node as a neighbour if there is an edge between both. Although it is a simple measure, it gives information about the importance of the node in the network. For unweighted networks, this value can be calculated according to the adjacency matrix:

$$k_v = \sum_{j=1}^{n} a_{vj}, \qquad 0 < k_v < n \tag{2.1}$$

where a_{vj} is the entry of the vth row and jth column of the adjacency matrix;
Or through the sum of all neighbours:

$$k_v = |N_v|, \qquad 0 < k_v < n \tag{2.2}$$

When direct networks are being used, the type of connection matters, which causes the degree to be calculated according to the number of incoming and outgoing connections. In a node v, the number of incoming links, also known as *in degree* k_v^+, is the sum of

connections that end in v. In contrast, the *out degree* k_v^-, is the sum of connections that begin at v.

$$k_v^+ = \sum_{j=1}^{n} a_{jv} \tag{2.3}$$

$$k_v^- = \sum_{j=1}^{n} a_{vj} \tag{2.4}$$

On a weighted network, the degree of a node is calculated through the sum of the different edge weights connected to v, according to the following expression:

$$k_v^w = \sum_{j=1}^{n} a_{vj^{\mathrm{w}}} \tag{2.5}$$

Another centrality metric is node **betweenness** b_v, which measures the number of times a node v acts as a bridge along the shortest path between two other nodes. For every pair of nodes s and t in a graph, exists at least one path between them such that either the number of edges or the sum of the edges weight is minimized. The betweenness of a node v, is the number of these shortest paths that include v dividing by the total number of paths between the two pair of nodes:

$$b_v = \sum_{s,t \in V(G)\setminus v} \frac{\sigma_{\mathrm{st}}(v)}{\sigma_{\mathrm{st}}} \tag{2.6}$$

Nodes with a high betweenness value (Fig. 2.6) are important actors on the network because they establish connections between different regions of the network.

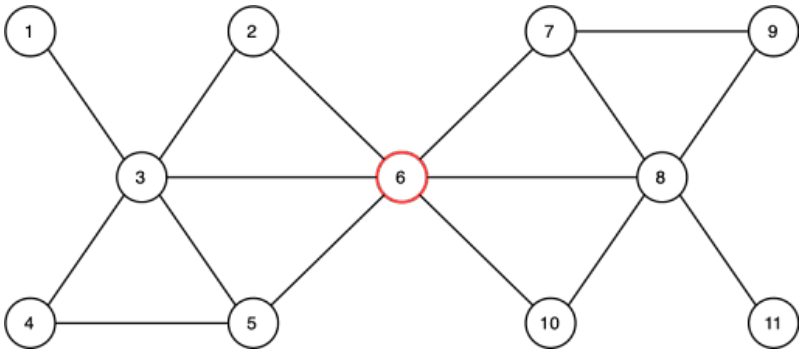

Figure 2.6: Node Betweenness Example

This principle can also be applied to edges in order to find the most important connections of a given network.

Closeness quantifies how well connected a node is to every other node in the network. Based on the mean length of all shortest paths between one node to all other nodes in the network, this measure can be calculated according to the following expression:

$$Cl_v = \frac{n-1}{\sum_u \in V(G)\setminus v \; d(u,v)} \tag{2.7}$$

where *d(u,v)* is the shortest path, or *geodesic distance* [9], between u and v. In the social networks context, closeness measures how fast a given actor can reach everyone in the network. The Figure 2.7 shows an example where the node marked with red colour corresponds to the element with the highest closeness value.

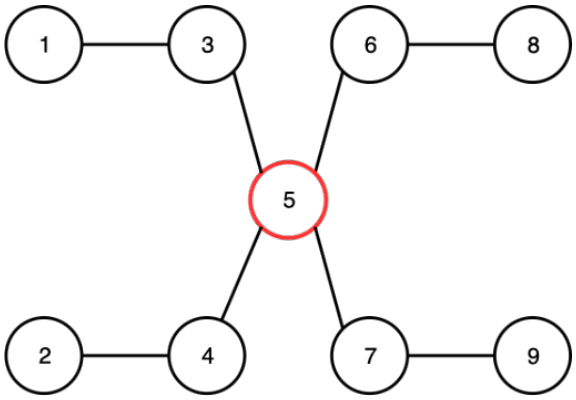

Figure 2.7: Node Closeness Example

Eigenvector centrality is a more elaborated version of the node degree. Relative scores are assigned to all nodes in the network and, based on this, node v is evaluated not only in terms of the quantity of connections but also in terms of their quality. These scores are assigned based on the first eigenvector in the adjacency matrix. This measure is based on the concept that the level of importance of a given node depends recursively on the level of importance of the nodes that are directly connected to it and so on. The formula representing this concept is defined as follows:

$$x_{\mathrm{v}} = \frac{1}{\lambda} \sum_{j=1}^{n} a_{\mathrm{vj}} x_{\mathrm{j}} \tag{2.8}$$

where a_{vj} represents the entry of the adjacency matrix, $\mathrm{x_v}$ and $\mathrm{x_j}$ refers to the centrality of node v and j, respectively and λ denotes the largest eigenvalue of the adjacency matrix.

The **Local Clustering Coefficient** c_v of a node v is given by the proportion of edges between the nodes within its neighborhood divided by the number of links that could possibly exist between them. This coefficient studies the level of transitivity in the node's neighbourhood indicating the level of cohesion between neighbors and can be calculated with:

$$c_{\mathrm{i}} = \frac{2|e_{\mathrm{jk}}|}{k_{\mathrm{i}}(k_{\mathrm{i}} - 1)}, \quad v_j, v_k \in N_i, \ e_{\mathrm{jk}} \in E \tag{2.9}$$

where N_i is the neighborhood of node v_i, E is the set of network edges, e_{jk} is the edge between the nodes v_j and v_k, $|\mathrm{e_{jk}}|$ is the proportion of edges between the nodes within N_i of node v_i and k_i is the degree of node v_i. This coefficient could also be computed for the whole network.

Network Level Measures

The **Diameter** and **Radius** are two network level measures that give us an idea of the distance between nodes. In order to understand these two it is important to define the

concept of eccentricity. **Eccentricity** of a node v is the greatest distance of the set of shortest paths between v and any other node. The Diameter d of a network is given by the maximum eccentricity of the set of vertices in the network. In the opposite way, the Radius is the minimum eccentricity of the set of vertices.

The **Average Geodesic Distance** is a measure of the efficiency of the information flow in the network. As the name suggests, it is the average of the geodesic distances between all pairs of nodes and can be calculated as follows:

$$l = \frac{1}{\frac{1}{2}n(n-1)} \sum_{i \geq j} d(i,j) \tag{2.10}$$

where *d(i,j)* is the geodesic distance between nodes i and j, and *1/2n(n-1)* is the number of possible edges in a network with n nodes. It is important to check that the network has no more than one connected component. In cases where it does, it is necessary to change the formula presented in 2.10 because for a pair of nodes where there is no path between them, the distance is defined as infinite. To avoid this problem, the **Harmonic Average Geodesic Distance** was defined:

$$l = \frac{1}{\frac{1}{2}n(n-1)} \sum_{i \geq j} \frac{1}{d(i,j)} \tag{2.11}$$

The **Average Degree** is another measure that evaluates the number of edges connected to a given node according to the total number of nodes in the network.

$$\overline{k} = \frac{1}{n} \sum_{i=1}^{n} k_{\mathrm{i}} \tag{2.12}$$

Reciprocity is a specific measure for the directed networks and quantifies the likelihood of two nodes being mutually connected. There are different ways to compute this metric but the simplest and easiest way is to calculate the ratio of the number of edges that point in both directions to the number of all connections:

$$r = \frac{\#rec}{\#rec + \#asym} \tag{2.13}$$

where *#rec* refers to the number of reciprocal edges and *#asym* is the number of asymmetric edges which are edges that only point in one direction for a pair of nodes. If *r=1*, the network is purely bidirectional. On the other hand, if *r=0*, the network is unidirectional. Generally, in real world networks, r varies in the interval between 0 and 1.

Network density is the number of edges divided by the total possible edges. High values of density ρ are associated with dense networks, and low values are associated with sparse networks. As in the case of reciprocity, the values of the density vary between 0 and 1.

$$\rho = \frac{m}{m_{\max}} \tag{2.14}$$

where m is the number of edges in the network and m_{max} is the number of possible edges, and can be obtained as:

$$m_{max} = \begin{cases} \frac{n(n-1)}{2}, \ for\ undirected\ networks \\ n(n-1), \ for\ directed\ networks \end{cases}$$

Global Clustering Coefficient is the global version of the Clustering Coefficient discussed above (Def. 2.9). This is obtained through the computation of the mean of all clustering coefficients of the nodes inside the network.

$$c = \frac{1}{n}\sum_{i=1}^{n} c_{\mathrm{i}} \tag{2.15}$$

K-shell Decomposition

The K-shell decomposition algorithm is a node ranking method that is used to identify the most influential nodes in a network [10]. It deconstructs the network into different levels (shells) based on the number of edges of a node. This algorithm assigns an index k_s to each node and this index represents the location of the node in the graph as shown in Figure 2.8. In the first iteration, all nodes with degree $k = 1$ are removed recursively from the network and a $k_s = 1$ value is assigned to them. This process is repeated iteratively until the network contains only nodes with a degree of $k > 1$. Then, all nodes with degree $k = 2$ are removed and the whole process is repeated. This method ends when all nodes in the network are assigned to one of the k-shell levels. Nodes in the core of the network are more relevant than the peripheral ones.

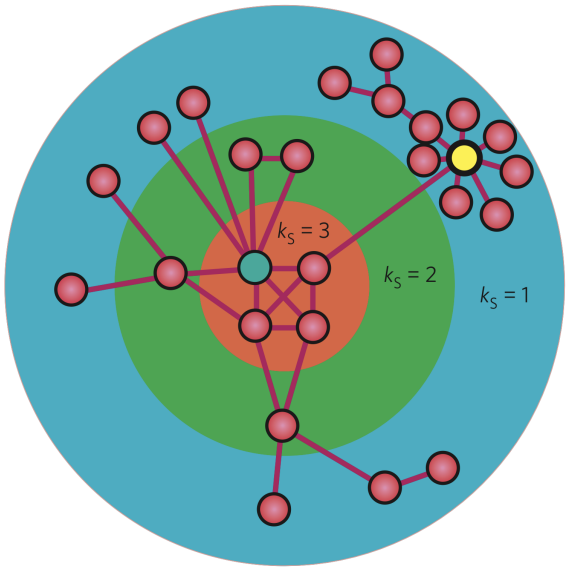

Figure 2.8: K-shell Decomposition Example Adapted From [1]

PageRank Algorithm

PageRank is a link analysis algorithm based on the concept of eigenvector centrality (Eq. 2.8). Presented by Page *et al.* [11] this algorithm was developed to measure the relative importance of web pages. Considering the web graph, where each node represents a web page and an edge a link between two pages, this method aims to measure the relative importance of nodes based not only on the quantity of links but also on the quality of these links.

According to the definition provided by Easley and Kleinberg [12], the algorithm starts by

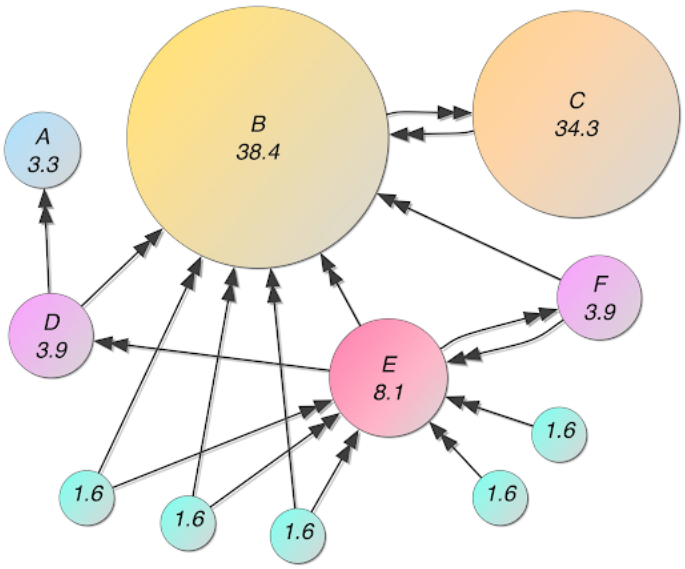

Figure 2.9: PageRank Representation

associating a relative score to each node of the graph. In a network with n nodes, this score is $1/n$. Then, the shares of all the nodes are calculated. These values are obtained by dividing the current PageRank score of node v by the number of outgoing links it has. If a node v has no outgoing links, the PageRank share is passed to itself. The PageRank score of v is then updated by summing the shares of the nodes pointing to it. This process is repeated until there is convergence in the PageRank values or after k iterations (being this value set initially).

Community Detection

People have the tendency to form groups. These groups can be families, friends, working groups, among others. They can be detected through the analysis of a network. Communities can be defined as densely connected groups of nodes in the network, with sparser connections between them [13]. In recent years, numerous algorithms have been developed for community detection. In this section we intend to highlight some of the most used algorithms and metrics that evaluate their output.

Girvan-Newman Algorithm

This is one of the most used algorithms and is based on a divisive hierarchical technique [13]. The algorithm consists of deconstructing the network into smaller connected components by removing the most important edges. In order to select this edges, the edge betweenness centrality measure is used. Based on edge betweenness, this measure finds the edges that lie on a large number of shortest paths between nodes. The steps of the algorithm are:

1. Computation of the centrality for all edges;
2. Removal of edge with largest centrality: in case of ties with other edges, one of them is picked at random;
3. Recalculation of centralities on the running graph;

4. Iteration of the cycle from step 2.

The algorithm ends when there are no edges remaining in the network. The edges connecting communities are then expected to have high edge betweenness. Generally, the output of this algorithm is an hierarchical structure, where it is possible to identify the communities at any level. This algorithm's time complexity is $\mathcal{O}(E^2V)$ [13].

Fastgreedy

This algorithm was proposed by Clauset *et al.* [14] and it is a greedy algorithm that aims to maximise the Modularity score. According to the authors, Modularity Q is a quality function that evaluates a certain division of the network into modules (Also called groups or communities). This measure values divisions where there are many links within communities and few between them. Modularity can be defined as follows:

$$Q = \frac{1}{2m}\sum[A_{\mathrm{ij}} - \frac{k_{\mathrm{i}}k_{\mathrm{j}}}{2m}]\delta(c_{\mathrm{i}}, c_{\mathrm{j}}), \tag{2.16}$$

where m is the number of edges, k_i and k_j represent the degree of nodes i and j, A_{ij} is the entry of the adjacency matrix that gives the number of edges between nodes i and j , $\frac{k_{\mathrm{i}}k_{\mathrm{j}}}{2m}$ represents the expected number of edges falling between those nodes, $\mathrm{c_i}$ and $\mathrm{c_j}$ denote the groups to which nodes i and j belong, and δ ($\mathrm{c_i}$,$\mathrm{c_j}$) represents the Kronecker delta which is 1 if $\mathrm{c_i} = \mathrm{c_j}$ and 0 otherwise.

The algorithm starts with the initial network, where each node forms a community. Then, the expected modularity improvement is computed for each pair of communities that can be merged together. The pair that gives the highest improvement in the modularity score is merged together. This process is repeated until there is no community pairs that increase the modularity value. This procedure can be represented as a tree whose leaves are the initial nodes and whose internal vertices correspond to the joins made. In this way, a Dendrogram (Figure 2.10) can be used to check the communities that have been merged in all levels.

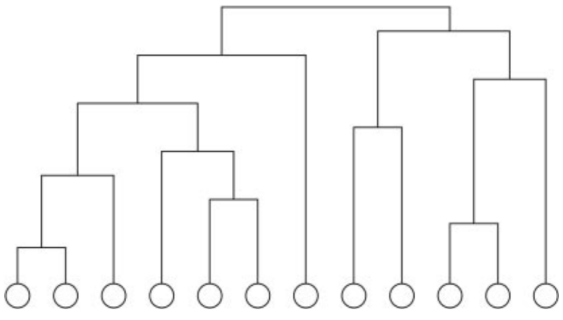

Figure 2.10: Dendrogram Example

It was found in practice that a modularity of about 0.3 is a good indicator of the existence of meaningful communities [1]. This algorithm runs in time $\mathcal{O}(MD\log V)$ for a network with V vertices and M edges where D is the depth of the dendogram. For sparce networks, the algorithm has a complexity of $\mathcal{O}(V\log^2 V)$ [14].

Label Propagation

This algorithm was introduced by Raghavan *et al.* [15]. Label propagation does not optimize any given objective function and it does not require to have *a priori* information about the network structure. In this algorithm, each node carries a label denoting the community to which they belong. Then, through an iterative process, the label of a node v is changed according to the most frequent label in their neighbourhood. In case of a tie, the label is chosen randomly. The process runs until every node in the network has a label to which the maximum number of its neighbors belongs. In the end, nodes with the same label are grouped together as one community. This is a widely used algorithm due to its near linear time complexity.

Leading Eigenvector

Similar to the *fastgreedy* algorithm, it aims to maximize the modularity score. Proposed by Newman [16], the main goal of this algorithm is the spectral optimization of modularity using the eigenvalues and eigenvectors of the modularity matrix B. Using the same notation in the Eq. 2.16, B can be computed as follows:

$$B_{\mathrm{ij}} = A_{\mathrm{i,j}} - \frac{k_{\mathrm{i}} k_{\mathrm{j}}}{2m} \tag{2.17}$$

Firstly the leading eigenvector of the modularity matrix is calculated, i.e., the eigenvector of the largest eigenvalue. Then, based on this eigenvector, the network is splitted in two parts such that the modularity improvement is maximised. Through an iterative process, the modularity contribution is calculated at each step in the subdivision of the network. This contribution is the difference between the modularities of the network before and after the subdivision step. The algorithm stops when this contribution value becomes non positive. Each graph bipartition has a $\mathcal{O}(V(M+V))$ computational complexity or $\mathcal{O}(V^2)$ for sparser networks [17].

WalkTrap

This algorithm assumes that communities are sections in the network with a higher density of edges and when random walks are implemented in the network, they tend to remain within these zones. Proposed by Pon & Latapy [18], this algorithm starts by assigning a community to each node and calculating all the distances between all adjacent nodes. Then, two adjacent communities are chosen and merged together through a hierarchical clustering technique. This selection is based on the minimization of the distances between the vertices inside each community. This process is repeated (V - 1) times and modularity can be used in the end, in order to select the best partition of the resulting dendrogram. Generally, the computational complexity of this algorithm is $\mathcal{O}(MV^2)$ but, for sparser networks, is $\mathcal{O}(V^2 \log(V))$ [17].

Louvain

This algorithm is a similar approach to the Fastgreedy method. Proposed by Blondel *et al.* [19], this method starts by assigning a different community to each node of the network. Then, for each node, the modularity change is calculated by removing it from its own

community and moving it to the community of each of its neighbours. The node is then placed into the community that resulted in the greatest modularity increase. If no increase is possible, the node remains in its original community. This process is applied repeatedly and sequentially to all nodes until no modularity increase can occur. A new network is generated where each community in the previous network is aggregated and replaced by a node. This process repeats until there is no change in communities structure or a maximum modularity is reached. The computational complexity of this algorithm is $\mathcal{O}(N \log N)$ [17].

Applications

With these measures and techniques, different ways of extracting information from the network can be developed. Some SNA applications include the development of recommendation systems, location-based interaction analysis, user attribute and behavior analysis, information propagation, among others [20].

Recommendation systems are mechanisms that we use everyday whenever we want to buy a book from a website, watch a movie or read an article on a social media site. These systems combine information based on the user's past behaviour (previously purchased or selected items), as well as similar decisions made by other users. These models are then used to predict items or products that the user may be interested in. Community detection methodologies can be applied to these systems in order to aggregate people with the same tastes. Fatemi *et al.* [21] proposed a Community Based Social Recommender System that produces a network of users' interests in movies and uses this as the basis for detecting communities and providing more accurate and personalised recommendations.

In the medical domain, community detection algorithms can also be used in order to discover the dynamics of the spread of a disease. Salathé and Jones [22] studied the impact of analysing the structure of a human contact network in order to provide knowledge to prevent the spread of diseases between communities. They have shown that in networks where a strong community structure is evident, it is possible to prevent the spread of disease between communities by delivering targeted immunisation interventions to individuals linking different communities.

The Influence Maximization Problem (IMP) is a topic that has been addressed by several researchers [23][24][25] and aims to identify a limited number of key individuals to spread influence in a social network such that the expected number of influenced individuals is maximized. By applying some of the centrality metrics discussed above, these individuals can be identified. From a business perspective, the identification of these individuals can have a big impact on how companies target their marketing campaigns. A company with a limited budget that wants to advertise its product in order to reach as many people as possible, can use the result of this selection in a way that these individuals can influence others to buy the product. In addition to this, community detection can also be used to directly segment their customers. Companies can provide better service solutions if they know their customer groups intimately.

SNA has been used in the detection of terrorist groups or other organized crime families. Based on call patterns, networks are established and analysed in order to find the group leaders. Special forces can plan attacks on these members, based on this data, in order to make the network more fragile [26]. On the other hand, Google uses the *Pagerank algorithm* to rank the web pages according to the information they carry. This algorithm is used as a mean to improve the search process in their search engines [27].

Due to technological advances, this type of technique tends to be increasingly used in

different areas. Due to the way data is organised, its analysis becomes simpler, producing results in a faster and more efficient way.

2.4 Related Work

Based on user presence patterns, different researchers developed different approaches to detect important locations to the mobile phone users [28, 29]. The most used are Home and Work (HW) locations. Some research works determine these locations based on the frequency of use during "Working Hours" and "Home Hours". Work hours and home hours are the time intervals where individuals are usually at work and at home, respectively. Despite being a good approach, other alternatives had to be studied because the time intervals vary from person to person.

Tongsinoot *et al.* [30], propose a method that identifies individuals' work places based on CDRs and Internet usage records considering the regularity and the time interval of use. On the other hand, they also present a method for the detection of home locations based on the use of the mobile phone during the rest hours of individuals. Based on the records of the last calls of the day and the first calls of the following day, the sleeping period is identified. With this time interval, the CDRs of the 4 hours before and the 4 hours after the calculated one are analyzed. The location is then selected based on the frequency of use.

Csáji *et al.* [3] analysed an anonymized communication dataset from a telecom operator in Portugal to prove that clustering techniques and principal component analysis allows a significant dimension reduction with limited loss of information. According to the authors, the features that provide more information are related to geographical locations. With this in mind, they cluster the different calling patterns and, for each user and each location, they show that there are only two types of locations that can be easily identified (HW). In the end, they identify these locations with a probabilistic inference framework and a maximum likelihood approach.

Leng [31] introduces in its research an additional feature to the CDR data called *Normalized Hourly Presence*. This feature captures how many times and how long each user appears at each location. According to the author, this allows not only identification of the frequency but also the temporal variations in each location on different days of the week. It is assumed that this feature is an indicator of the type of location for each individual. Then, using two clustering techniques (K-means and Fuzzy K-means) home and work locations are identified. In the same document, Leng presents other research works using CDRs. Through cell tower traces for each individual, the author proposes the use of Recurrent Neural Networks in order to predict the individuals next location and it is proven that CDR data is useful to make location recommendations.

Vanhoof *et al.* [32] investigated and compared the use of five different criteria for home location detection. With the introduction of a measure of spatial uncertainty, they outperformed the simple home detection algorithms described by the literature. This uncertainty is calculated based on three different locations returned by the algorithms, where Location 1 is the home location and Location 2 and Location 3 are the other plausible home locations. The value is then determined regarding the distance and the number of observations of the criteria in each of the three locations.

Other works have been published regarding the movement patterns of individuals [33, 34, 35]. Due to daily life activities, human movements have a high level of regularity. Regarding

this, some researchers have decided to use the CDRs in order to study the mobility patterns of individuals and provide information about the structure of cities and their dynamical properties. Louail *et al.* [36] showed in their study that it is possible to obtain meaningful information from mobile phone data, regarding not only the mobility of people, but also on the structure of the city itself. They started by defining an urban dilatation index which measures how the average distance between individuals evolves during the day and then, they proposed a method to determine hot spots, which are the most crowded places in the city. With this information, it is possible to analyze the distribution of the population during different hours of the day.

As mentioned above, CDRs are not only used for the detection of important locations, but they also allow the study of relationships between individuals. To do so, some researchers used network analysis techniques in order to establish relations between individuals or communities [37]. Al-Molhem *et al.*[1] published a paper in which it is applied SNA concepts to model their social network in a weighted graph based on call records. In this structure, each relation carries a different weight representing how close two nodes are to each other. Then, based on centrality measures, they apply social network analysis concepts in order to determine the node importance. This measure is determined by the level of importance of the node neighbours as well as number of nodes linked to this node. The nodes with the highest values are considered the *influencers* of the network.

Rum et al. [5] propose an alternative to calculating network influencers based also on centrality metrics. The social networks were built with data collected from the twitter API, that allows public users to access and extract huge amounts of data. Then, five centrality measures (in degree, out degree, betweenness, closeness and eigenvector centrality) were applied to these networks in order to find the most important users. In the end, the researchers came to the conclusion that the most important users in the network are those who are in a more central position and that their level of importance depends a lot on the other actors to which they are connected.

Churn, also known as Churn rate, quantifies the number of customers a company loses in a given period of time. Ngonmang and Viennet [38] used data from a social platform called Skyrock in order to analyze and predict the propensity of a user to stop using a social network platform in the near future. They start by detecting communities in the corresponding network graph using two algorithms: Louvain's and IOLoCo algorithm [39]. Based on the resulting communities, attributes are extracted from each of the nodes such as degree, size of the local community, number of elements in the same community that have been inactive for some time on the platform, among others. These attributes are then used to train standard Support Vector Machines that support the classification of each of the nodes in the network. Through the analysis of the results obtained, the researchers came to the conclusion that the attribute that most contributes to the prediction of churn is related to the proportion of inactive elements in the same community strengthening the hypothesis that individuals are strongly influenced by the behaviour of other community members.

Nathan Eagle *et al.* [40], compared data collected from mobile phones of individuals with their own self-report data. In addition to call logs, researchers also had access to nearby Bluetooth devices, application usage history and cell phone status. Based on this collected data, and using a method called *Nonparametric multiple regression quadratic assignment procedure*, they proved that it is possible to infer 95% of the relationships between individuals. Moreover, they compared the information from both data sources relative to the time two people spend near each other and the type of relationships between them, suggesting that there are differences between the two when, for example, individuals

try to remember recent interactions.

Wu [41] developed a similar study to the previous one, but in this case, it was studied the data extracted from the CDRs in order to identify the family relationship between individuals. According to the author, the duration of a call between two family members can vary from 25 seconds and 75 seconds. He started by extracting and identifying the individual's calling patterns and then, this data was used by machine learning models to predict the family relation of an individual. Several techniques are compared in order to obtain the best classification accuracy. In the end, the author concluded that it is possible to extract family relationship classification models of the individuals calling records.

2.5 General Data Protection Regulation

When confronted with the use of real data, whether personal or collected through the use of different tools by people in their daily lives, companies have to be careful in the way they manage and store this data in order to respect the rules imposed by the General Data Protection Regulation (GDPR).

The GDPR is a European regulation that took effect on 25 May 2018 and establishes rules concerning the processing, by a person, a company or an organisation, of personal data relating to individuals belonging to the European Union [42]. In other words, it aims to give citizens ways to control how their personal data is handled by third parties.

This regulation had to be introduced to address the challenges raised by the technological evolution that has taken place in recent decades. The amount of data stored and exchanged, whether structured or not, across borders has increased and will continue to increase, forcing the introduction of a more robust legal framework.

These rules have increased people's rights. They can access at any time the information that the entity holds about them, and they have the right to choose whether they want their data to be processed, for how long and for what purpose. In short, organisations can only manage this type of data with the customer's consent. This represented a major impact for companies that had to restructure several internal processes to comply with the legislation.

To comply with the GDPR, companies had to understand what kind of information they held, how they used it, where and how they stored it and which systems or organisational models had to be changed. In addition, companies whose business focuses on the regular processing of personal data were required to appoint a data protection officer (DPO), whose mission is to ensure compliance with all the rules and declare any problems that it may identify, suggesting ways to solve them and also getting involved in their resolution.

Despite these impacts that the GDPR has brought to companies, they have not stopped processing personal data, some because the processing of this type of data is necessary for their operation and others because this data helps improve the quality of their services. People's actual usage patterns are frequently used to understand the preferences and needs that individuals have in their day-to-day tasks. With this information, companies can adapt the way they operate to meet the identified needs.

To comply with the rules applied, many companies rely on different techniques that allow data to be used without compromising people's personal data. One of these techniques is the anonymisation of fields where it is possible to identify customers. Data anonymization refers to the method of preserving private or confidential information by deleting or en-

coding identifiers that link individuals to the stored data, preserving the credibility of the data collected [43]. There are different anonymisation techniques which are summarised below:

- **Data Masking**, as the name suggests, refers to changing the data with modified values. Generally, when this technique is used, a mirror image of the database is created and modification techniques such as character shuffling, encryption or even character substitution are applied.
- **Pseudonymization** allows private identifiers to be replaced by false identifiers, for example a person's name to be replaced by another. This technique allows privacy to be respected and data integrity to be maintained.
- **Generalization** is a transformation where some attributes are removed to make the data less recognisable. The aim is to remove some of the identifiers while maintaining the accuracy of the data.
- **Data swapping** is as simple as it seems and implies swapping the values of different attributes in the dataset in different columns so that they do not fit the original information.
- **Data perturbation** is applicable to numerical data entries. This technique rounds the values of these fields with a specific value or operation. (For example: multiply the Age field by 3)

When the data to be used can not be anonymised, some machine learning techniques can be used in order to create synthetic datasets. Synthetic datasets are datasets created by an algorithm without any connection to real events [44, 45]. They are built from the original datasets intended to be protected and, sometimes the results produced, may meet specific needs or certain conditions that may not be found in the original, real data.

By using these techniques, companies can overcome the barriers imposed by the GDPR on the use of personal data. In addition, they can store data in a more secure way by adding an extra layer of security. In the case of the existence of a database breach, where possible attacks could be made, the datasets would still seem incomprehensible if they were perfectly anonymous.

2.6 Discussion

Throughout the literature and to the best of our knowledge, no works have been found that relate directly to what we intend to achieve. However, some of the techniques discussed above may be important for this work.

As discussed, Call Detail Records monitor the mobile activity of individuals. Due to the relational nature of this data, networks of interaction can be built. These networks can then be studied using social network analysis techniques in order to understand how individuals are related. For example, the metrics described above are important in the sense that they not only allow us to understand the networks and the way they are structured, but also allow us to study the importance of the nodes that constitute them, helping to understand the relationship they have with each other and study their influence at the network level.

On the other hand, community detection algorithms are important when it comes to detecting relationship groups. The identification of communities in general can be useful to

understand communities at the level of networks and relate them in real-life relationships, understanding the relationships between individuals and how they are connected. It is of course important to mention that the detected groups can be heterogeneous with respect, for example, to their size. This and other factors have to be taken into consideration during the development of the models.

The analysis of call patterns and their locations could be an important point in the research as elements of the same household share the same relative home location. Using the results of this analysis may be important from the point of view of validating our approach.

It is important to note that the CDRs have some drawbacks. These records are event-driven, which means that the record is only stored when a connection such as a phone call, text message or web browsing, is made. Infrequent mobile users may not be considered by some of the methods presented which can lead to bad classification results. On the other hand, call records have a low spatial resolution. The coordinates of the cell towers approximate the geographic location of the individuals but the detection of the true location generates a big challenge especially in dense and compact areas.

The use of this type of data may also imply GDPR violations. Although this data does not provide the content of the communications established between individuals, it allows for example through the locations of the towers used to obtain a nearby location and track the movement of customers over a certain period of time. Companies collecting and managing data of this nature must always be careful to respect the rules imposed so that they do not violate the law and jeopardise their business. For this, it is necessary that customers are warned and give permission for the management of this type of data to be done in a legal way.

We consider that the use of this type of data is always beneficial when applied to the improvement of a certain product or service. However, it is always necessary to make an assessment in order to verify if the information that is intended to be used allows us to identify people. When this happens, we believe that the use of some of the anonymisation techniques discussed above can be useful in overcoming these challenges.

In the end, the use of this data is advantageous not only for the companies because it allows them to increase their business volume but also for the customers because when they receive tailored products, it makes them feel better and improves their life quality.

Chapter 3

Methodology

This chapter provides a more detailed analysis of the problem at hand. Based on the CRISP-DM methodology, the different phases that constitute this methodology are presented, relating each of them to our work. We begin by exploring the objectives of this work by giving the reader a contextualisation of the problem. Next, we describe the dataset and the validation process carried out. Furthermore, the exploratory analysis of the data performed is presented, which allows us to extract the main characteristics of the available data. At the end, the transformations that had to be done on these data in order to make them ready to be used are also presented.

3.1 Business Understanding

Telecommunication services have expanded dramatically due to increasing demand among consumers, as many companies and private individuals need these services to help their business expansion and daily lives. However, due to the competitive nature of the market, there is an increasing pressure being put on telecom service providers. Due to this pressure, companies have to ensure that they are competitive in the market segments in which they operate and also guarantee the satisfaction of the customers they have in order to keep or secure new ones. According to Kotler and Fox [46], winning new customers costs between 5 and 7 times more than keeping the same customers that a company already has.

In an attempt to mitigate this problem and establish a good relationship with their customers, companies are increasingly investing in a technology called Customer Relationship Management (CRM) [47] which, as the name suggests, aims to help companies to be connected with their customers to improve their commercial relationships. CRM systems start by collecting customer data from different sources and channels. Then, based on this information, they enable companies to learn more about their target audiences and how best to meet their needs, thereby retaining customers and driving sales growth. With a consolidated view of each customer, a CRM system is then used to manage day-to-day activities and interactions with customers. From a marketing perspective, this means engaging a company's customers or prospects with the right message at the right time through targeted marketing campaigns.

Companies, especially those in the telecommunications sector, have customer pools that are made up of millions of people who, despite sharing the company's services, are heterogeneous, with habits, tastes and demands that differ from one another. This situation requires specialized marketing methodologies for an effective performance with the target

audience.

One way to overcome this is to create profiles of customers through segmentation techniques. Market segmentation involves the division of a large heterogeneous set of customers into clearly identifiable segments. Customers are segmented based on whether they meet a particular criteria or share common interests, desires or expectations, causing them to have similar product requirements. Generally, these segments are made up of customers who will respond similarly to marketing strategies and campaigns. Having clear defined segments allows companies to satisfy a variety of customer needs, offering specific packages and incentives contributing not only to their fidelization but also to their retention.

With this in mind, we have as our main objective with this work the segmentation of customers according to their household and social relationships. Starting from the hypothesis that individuals belonging to the same household establish a proximity relationship that can be confirmed through their mobile traffic, we intend, through the study of Call Detail Records collected by a national telecommunications company, to use Social Network Analysis techniques to infer the existence of this type of groups. Additionally, we also intend to infer the distributions of telecom operators, within the detected groups.

In addition, we aim to develop a model to calculate the level of influence of each individual in the network. Through the use of some network metrics we aim to infer the elements with the highest level of centrality, since from the marketing point of view, a company with this type of information can run specific campaigns for these individuals since they can influence others to buy a particular product. On the other hand, within a household, there is the possibility that the individual who has a higher centrality score is considered the decision maker. With this information, companies can also adapt the way they build campaigns and approach the customer to improve their success rate.

Through the creation of these models, we pretend to improve the knowledge that a telecommunications company has about its customers not only to increase its productivity levels but also to improve the way it approaches its customers and the relationship it has with them.

Despite the results produced by the models, we also have the goal of optimising its computational performance so that it can support a large amount of data. Due to the nature of the problem, large amounts of data are produced daily which, when analysed at national level, require appropriate models in order to obtain results in a time-efficient manner.

3.2 Data Understanding

Dataset

The source data was provided by a national telecommunications company that is associated with the development of this work. The data was collected between September and October 2020 and contains Call Detail Records (CDRs) from calls, SMS and internet usage of customers in the city of Coimbra. To better understand the main dataset structure, the scheme shown in the Figure 3.1 can be introduced.

The Call Detail Records data is divided into 3 pairs of tables. Each pair relates to a type of communication, i.e., voice calls, text messages and internet usage (Also called General Packet Radio Service (GPRS)). Inside each one of these, the data is divided into two types of mobile contracts. The ones with the *PPP* terminology refer to the communication records of the Post Paid Plan (PPP) clients. Those who do not have this terminology, are

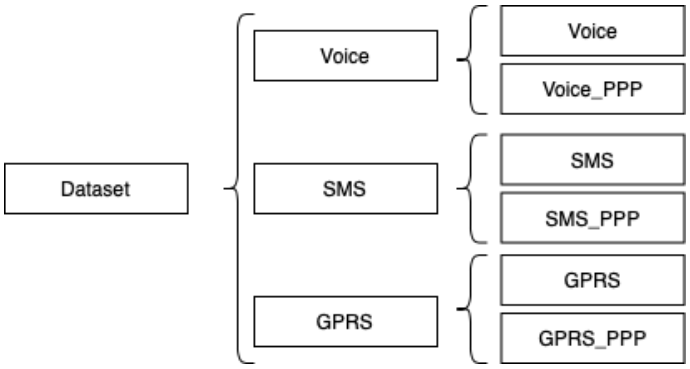

Figure 3.1: Diagram Of The Main Dataset Tables

related to the Pre-Paid clients.

The number of records in each table varies between 335014 and 4996508 and, for each one, it is possible to consult when the transaction was made with a level of precision to the second, the number and code of the mobile card involved, the mobile phone tower and two other codes that characterise the customer. In addition to these, there are specific features of the type of operation we are looking at. If we are looking at the voice call tables, we can assess the number to which the call was made i.e., the destination, the duration, the mobile operator and the type of call. By analysing the latter, we can check, for example, whether the call was made abroad or vice versa. The SMS tables have the same features as the voice tables, except for the duration, transmitting the same type of information. On the other hand, if we are analyzing the tables regarding the use of GPRS data, in addition to some features that have already been mentioned, it is possible to consult the amount of kilobytes spent and the duration of the GPRS session.

Additionally, a table with customer information was also provided. There are 33547 clients with 8 features about account details such as Client Identification Number (NIC), activation account date, deactivation account date, telephone's number and code's market, which indicates to which market the customer belongs (particular or PME). In addition, we have access to information about the cards associated with each customer. There are 43904 records with 8 features about the card by itself, like the card's code and telephone's number, but also, it contains information about the client to which the card is linked, such as NIC, career's code, number of children, age and sex are personal features, which does not correspond to the card.

Two tables regarding interactions and solicitations between the CRM and customers were also supplied. The interactions table contains 108088 interactions described by 16 attributes. In general, the information carried by it is about the date when the interaction was initiated, the date when the interaction was finished, the channel used, the direction of the interaction (inbound or outbound), which client is related to the interaction and a solicitation identification number. The solicitations table has 45355 records described by 20 attributes and contains information about the context of the solicitations, its start and end date, direction (inbound or outbound), channel used and customer identification fields.

To complement these main tables, reference tables were provided to understand the meaning of each code. For example, the cell tower reference table allows us to check, for each tower, its exact location both in terms of GPS coordinates and the street address.

It is important to note that a selection process of the most important features was per-

formed and only the records of private customers have been selected, discarding companies and other non-private entities. This selection process was discussed and analysed in a meeting with all the elements linked to the project. On the other hand, fields such as phone numbers and card numbers were anonymised to comply with data protection policies. This anonymisation process was done by the company that provided the dataset and the values of these fields encrypted.

Data Model and Database

When dealing with huge amounts of data, we need to ensure its integrity. Each of the files received contained data that were related and complemented each other. As a consequence, we decided to create a relational database to organise all the information in a structured way where the relationships between them were explicit. Our main concern was to maintain the correct connection between the data to avoid changing the meaning of the information during the whole process. To do this, we decided to create a data model that allowed us to explicitly represent the structure of the data

The data comprehension phase underwent many iterations, as many questions arose about the meaning of some characteristics and some values. Since the data model definition was dependent on the outcome of this phase, the data model had to be rebuilt a few times having some impact on the work's progress. The process of understanding the data, creating the model and building the database was done in collaboration with other research work that, although focusing on different analysis, uses the same data source.

The final data model is presented in Appendix A - Fig 1. Although all the data have been processed and organised, only part of the tables represented in the model have been used in this work. The segment of tables relevant to this work is presented in Appendix A - Fig 2.

The database was implemented in PostgreSQL [48] and a SQL script was developed that creates each table with the relations that each one has. The choice for PostgreSQL was due to the fact that we are more familiar with the tool, thus reducing the learning curve.

Data Validation

After we become more familiar with the data provided and in order to ensure its consistency, a data validation process has been done. Data validation ensures that the data is complete, unique and the range of values is consistent with what we expect. This is an important step because it prevents decision making on data that is not accurately representative of the situation in study. It is important to mention that this process was only applied to data we considered relevant for this work. Data related to interactions and requests with the CRM were not validated since they are not related to the objective we wanted to achieve.

This process can be divided into three main operations:

- Check for missing values;
- Check for outliers;
- Check for malformed fields;

As a consequence of using a PostgreSQL database, a portion of these operations was made automatically when the data was loaded. During the creation of the database tables, we

needed to define the type and the non null attributes. By defining all the attributes of the table as not null, we can ensure that, if the loading is done without any error, there are no missing values in the tables. During this process we came to the conclusion that customer related demographic features such as age, number of children contained a large amount of null values. For the remaining data, as we were able to do the loading without any problems, we considered that there are no attributes with null values and the data type corresponded to what we expected.

Regarding the outliers, we only needed to check the time related attributes. The main idea was to check whether they were within the originally defined interval (September and October 2020). For this, a small script was developed that went through all the records of the tables that contained these fields in order to perform this verification. Attributes such as cell tower code and customer characterization codes have an associated reference table. Therefore, since we defined these attributes as a foreign key for another table, PostgreSQL ensures that the value is present in the reference table. In general, no outliers were found in the data.

It should be mentioned that in the cell tower reference table, we found that there were a small number of entries with the same attribute values. This happens when the real location of the cell tower is not defined. After analysing in more detail, we decided not to do any transformation on this data as we found that the cell towers defined in this value range were not used in our dataset.

3.3 Exploratory Data Analysis

Before the modeling phase, we decided to perform an exploratory data analysis on the dataset provided in order to study its main characteristics. We chose to use only the voice call records not only because it is the most widely used throughout the literature but also because we consider that this type of interaction translates into a higher level of proximity giving more information about the relationships between individuals.

As mentioned above, our dataset contains two tables for the voice calls, each one referring to the type of mobile contract. The following table shows the number of entries for each.

Trafego Voz Pre Paid	Trafego Voz Post Paid
335013	3803400

Table 3.1: Number of Dataset Entries

As we can observe, both tables contain a large number of records. In general, the tables for customers with post paid plans (PPP) have a higher number of records than the remaining.

Primarily, we started by studying the time distribution of data, i.e., to find out if there are different patterns in different timelines. To simplify the analysis and the visualization, we developed a dashboard using two Python modules: Plotly and Dash [49].

With weekly time intervals, we started by checking the days and hours with the highest traffic. The Figures 3.2 and 3.3 illustrate the traffic heatmap in the second and third week of September. The colours represent the number of calls made at each hour of the day.

In general, the interval between 10 am to 8 pm has the most traffic, corresponding to 82% of total calls. On the other hand, the days when people make the most phone calls are Monday and Friday afternoon, with a big break at weekends (17%). These results were

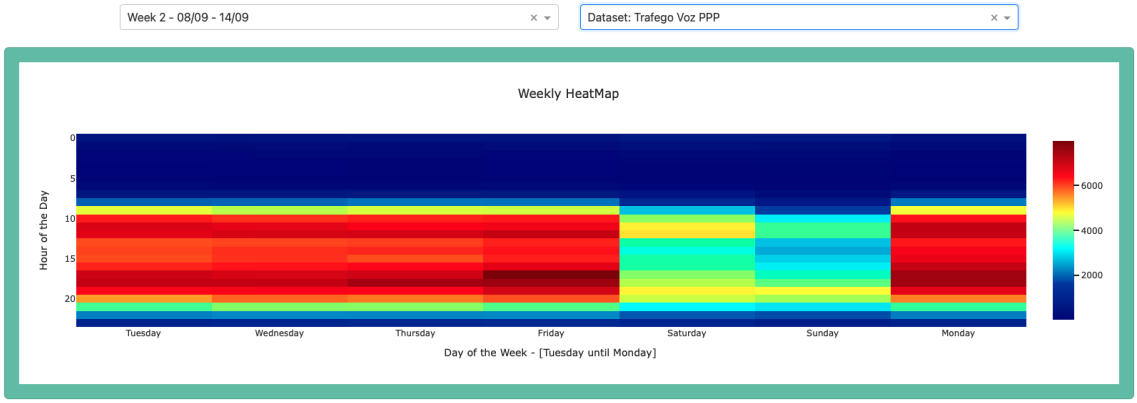

Figure 3.2: Weekly Heatmap - Second Week September

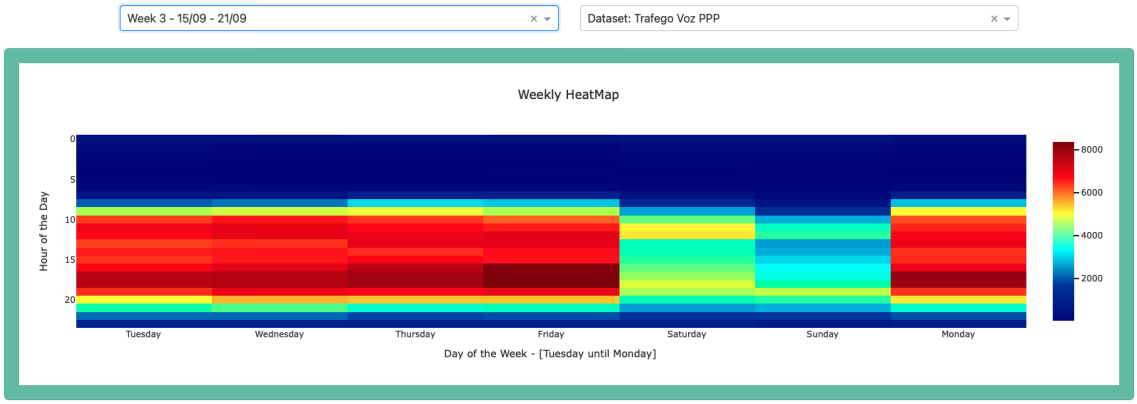

Figure 3.3: Weekly Heatmap - Third Week September

also discussed with the company team involved in this project where they confirmed that these results are in agreement with the records they take from the company.

Then, to understand this distribution by number we developed a similar heatmap where we select the number we want to check. This selection can be made of an increasing or decreasing order of the number of calls received or made on the selected week. Figure 3.4 shows the heatmap for the number 347188.

Through this graph, we analysed the presence of calling patterns within different weeks. Although this is a direct method, no patterns were found in the numbers analysed. Besides, it is important to mention that during the analysis of this graph some abnormal values were detected. When selecting the numbers with the highest number of calls, we found that those that came first received over 600 calls per week. As the records are related only to private customers, we consider this to be a high value. In total, six numbers with these characteristics were detected. To better understand this, a call duration filtering was implemented. We found that the call times for some of these numbers are always zero but, for others, this is not the case and there is no clear distribution over time. The figure 3.5 shows an example.

To understand these results, we send the anonymised codes associated with these numbers to the the company's team to evaluate this data. After analysing these numbers, we were informed that, despite these results, the customers associated with these codes do not have any characteristic that differentiates them from the others. For this reason, we decided

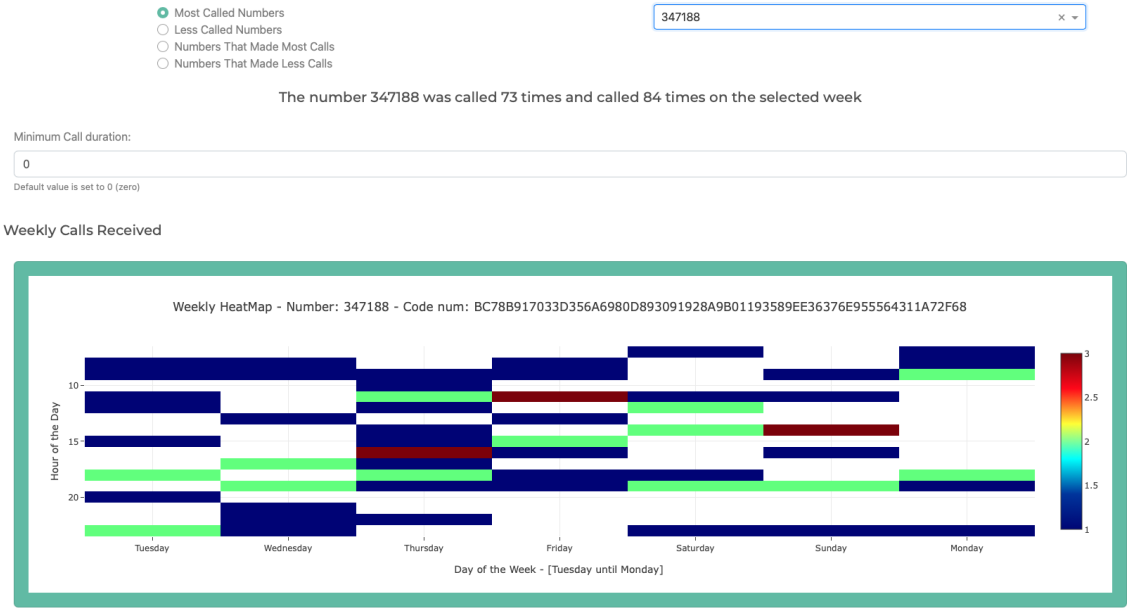

Figure 3.4: Weekly Heatmap - Number 347188

to keep these numbers in our dataset always bearing in mind that these outliers can have effects on the results obtained by the models.

Finally, to inspect the recurrence of calls between different numbers, we analysed the number of inbound and outbound calls between pairs of numbers. An inbound or outbound call may also be called an incoming or outgoing call, respectively. Figure 3.6 shows an example of the recurrency of calls in the first week of October for the number 420503. This representation also allows us to verify whether a given number only receives or makes calls.

In this example, we can observe some call recurrence between the selected number and three other numbers. Assuming that when there is recurrence between numbers, their level of proximity is higher, this group of numbers may represent a household or a social group.

Through the analysis of this type of graph, we were able to see that there is some recurrence of calls between pairs of numbers over the different weeks. Because of the amount of different numbers, it was not possible to analyse all of them, but the information we have drawn from the analysis of these graphs already allows us to reach the first conclusions:

- There is more traffic on weekdays and less on weekends.
- The time interval where the most calls are made is between 10am and 8pm (82%).
- There are a large amount of numbers that only make or receive calls.
- There is some recurrence of calls between numbers but this is not very perceptible with weekly divisions. The fact that there is some recurrence could be important to our problem. This tells us that there are customers who establish contact with each other which may indicate that exists some kind of proximity between them.
- No call patterns were detected during the different weeks for the set of numbers analysed.

Figure 3.5: Weekly Heatmap - Number 78220

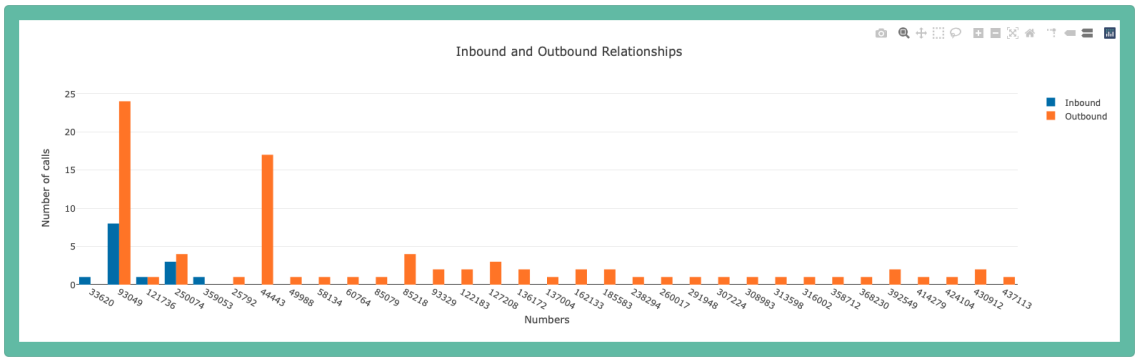

Figure 3.6: Interactions Graph - Number 420503

- There is a low number of calls where the duration is zero taking into account all the records in the tables (0.4% of total calls). This shows that the calls made really do represent an interaction between individuals. If the percentage of calls with zero duration were higher, we would not be able to assess whether a given pair of individuals actually know each other.

- For the post paid voice dataset, there are no records of calls that have been made in the last two weeks of October.

3.4 From Operators to Analytics

The raw data must be processed before it can be used. It is often necessary to carry out a process of selecting, cleaning and transforming the data so that it can be used in the modeling phase. When the data was received, some operations had to be done to make its representation simpler and easier to understand and process, always respecting the defined data model. This was a time-consuming process due to the large amount of data available.

These operations were done mainly on data that has been anonymised. Fields such as phone numbers and card numbers have been anonymised and their value has been replaced with an alphanumeric string. To improve this representation, not only to facilitate the analysis of results and database queries but also to reduce the cost of processing, two reference tables had to be created, one for telephone numbers and one for card numbers. An integer has been associated to each alphanumeric value in order to create a unique identifier. According to this association, the alphanumeric values in the main tables were then replaced by their unique identifiers.

Regarding the call records tables, we decided to end the distinction between prepaid and postpaid customers (PPP), merging both tables. We consider that this distinction is not relevant for this work and by joining both tables, it makes the database queries simpler.

Then, it was necessary to change client related fields such as sex, age, number of children and career from the cards table to the client table since we consider that it makes more sense if a customer possesses such attributes. In the course of this process some customer codes were discovered that had no match in the client table. As a consequence, we decided to add them to the customers' table with the remaining attributes empty.

Depending on the data codification used to write the files and the type of reading we were making, some values had a mistaken type. One example was the field age from the table client, which was read as a float. To overcome this, a Python script was developed that allowed changing the type of variables of a given column.

This page is intentionally left blank.

Chapter 4

Influencer Detection

The identification of the most influential individuals (also called Influencers) along the customer base of a given company is important when it comes to planning and creating new campaigns. Companies can target these individuals for the purpose of supporting their brand, product or service, and in some cases, driving their purchases.

This chapter aims to describe all the steps and decisions taken in each phase of the influencer detection model development. In addition, a way to infer the mobile operator associated with each phone number is presented. The results obtained were analysed using a visual tool developed specifically for our work.

4.1 Classification Model

The construction of a model to identify the influencers in a network consists of three main steps:

- Building interaction networks;
- Calculation of centrality measures;
- Ranking clients by influence;

Each of these steps will be analysed in detail below so that the reader can understand the details behind the development process.

Building Interaction Networks

The construction of the interaction networks is a crucial activity in the development of this work. Through the analysis of these data structures, we can study the relationships between individuals and how the flow of information occurs.

As already mentioned, our dataset consists of data collected in the months of September and October 2020. Due to the high amount of data, we decided to build weekly networks to make the whole process more efficient. The weekly divisions are shown in the Table 4.1. It is important to mention that the classification model is independent of the time interval defined initially. This means that the model accepts interaction networks with any time interval.

Week 1	01/09 - 07/09
Week 2	08/09 - 14/09
Week 3	15/09 - 21/09
Week 4	22/09 - 28/09
Week 5	29/09 - 05/10
Week 6	06/10 - 12/10
Week 7	13/10 - 19/10
Week 8	20/10 - 26/10
Week 9	27/10 - 31/10

Table 4.1: Weekly Divisions

For each week, voice call detail records were grouped by number according to the following structure:

Caller Number | Called Number | Number of calls

where the caller number is the one initiating the call and the called number is the one that receives it. All these data filtering and grouping operations were done through an SQL query during the generation of the networks. Each entry in the result set corresponds to an interaction between two individuals. With this data, a network is built with the help of the python module Networkx [50]. This module provides a set of tools that allow us to build, manipulate and study the structure of complex networks. From a structural perspective, each number corresponds to a node in the graph and the number of calls is used to define the weight of the edge between the two nodes. In addition to being weighted, the networks produced are also directed, since the direction in which calls are made is important for us to understand how the flow of information in the network occurs. For example, nodes that establish connections in both directions between themselves may be clients that establish a relationship of greater proximity relative to those where this behaviour is not observed. After the network is created, it is stored in the database. All the tables created and their attributes are described later in this section.

Calculation of Centrality Measures

Influencer detection is based on four state of the art centrality measures: In degree and Out degree, Betweenness, Page Rank and Closeness. These measures were analysed in the centrality measures section (Sec. 2.3). Centrality measures were computed for each network built in the previous step and the results were stored in the database so that they are always available. As with the network construction, methods from the Networkx package were also used to compute these metrics, and the results are returned in dictionary format with all values normalised.

It is important to mention that for the calculation of the betweenness centrality the inverse weight of the links was used. The Networkx method that calculates the betweenness centrality uses the Dijkstra algorithm to obtain the shortest path between a pair of nodes. This algorithm always values the links with a lower weight so that the chosen path has the shortest possible distance. However, in our case, we want to value the links with the highest weight since it is through these that, theoretically, a greater flow of information occurs.

Also, only 30% of the network nodes were used in the shortest paths calculation. The difference in values is barely perceptible and the processing time decreases considerably.

The calculation is done offline due to the complexity of some of the metrics used in the model. Due to the large amount of data, some of the metrics take some time to process, which means that the results cannot be computed in real time. By doing this processing in offline mode, the results are always available when they are needed.

Ranking Clients by Influence

Based on the results of the metrics calculated in the previous activity, the level of influence of each customer is calculated with the following expression:

$$\textbf{Influence level} = \frac{InDegree + OutDegree}{2} + Betweenness + PageRank + Closeness \tag{4.1}$$

Customers with the highest influence values are the influencers in the network. This expression is only valid for customers who have a non-null betweenness value. As already discussed above, during the analysis of the dataset provided, a set of numbers that only received or made a high number of calls were detected. Due to the high degree value, most of these customers were being considered as network influencers. In order to avoid this situation, we consider that the level of customer influence is zero whenever the value of betweenness is also zero, since this metric tells us which customers are information brokers.

As with centrality measures, the level of influence for each customer is also stored in the database so that it is always available.

4.2 Client Classification by Mobile Operator

Identifying the mobile operator of the customers becomes an important factor from the point of view of a telecommunications company. From a marketing point of view, organisations, by identifying individuals belonging to other mobile providers, are able to develop targeted campaigns trying to persuade these individuals to change their mobile operator.

Call logs record to which mobile operator a particular call was made. With this information we are able to associate the operator to a certain number as long as it has received a call. So, for each number in our database, we fetched the last call it received and identified the mobile operator code associated with the register. This was done through an SQL query that starts by fetching all the calls that a given number has received and puts them into a temporary result set where they are sorted in descending order of the day on which they occurred. Then, only the first row that corresponds to the last call that number received is returned.

There was a concern to retrieve the last record in order to avoid situations where a particular customer changes operator during the time interval of the data. An assessment was made in order to confirm if this hypothesis was true. This verification was also done through an SQL query and we found that there are 65 numbers that changed their mobile operator. For customers where we have no record of incoming calls, the mobile operator is classified as undefined. This data was all stored in the database for future use.

4.3 Visualisation Tool

The analysis and validation of results is an important phase in the development of a classification model because it allows us to check whether the results obtained are in accordance with what we expect. Due to the nature of our data, an analysis of the results obtained through database queries would be a cumbersome and unclear operation. Thus, we decided to develop a dashboard to perform this analysis in an interactive and faster way. This dashboard was developed in Python using the Plotly Dash module [49]. Figure 4.1 shows the network for the first week of September with K-Shell level 6.

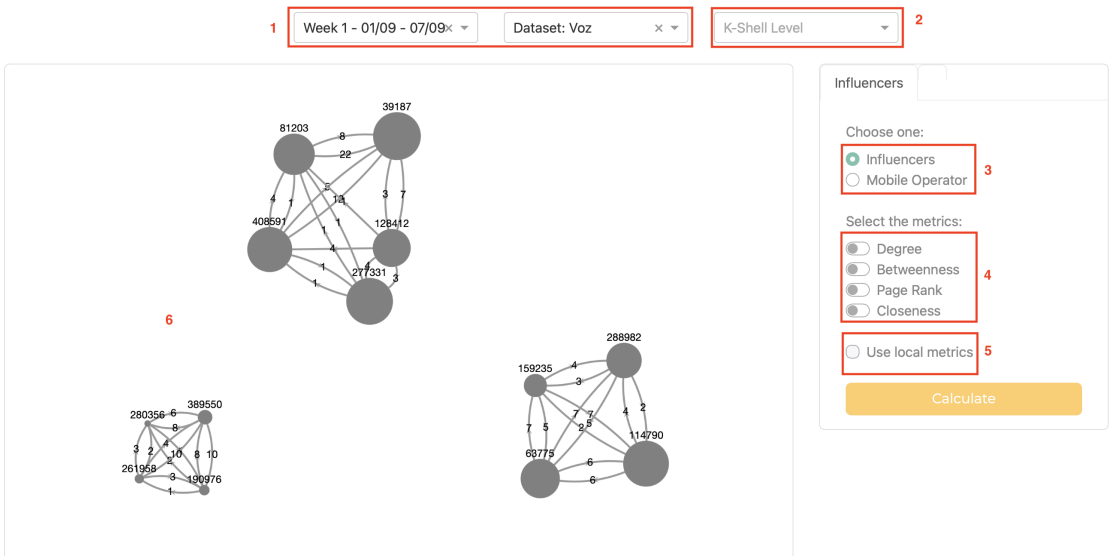

Figure 4.1: Preview of the Developed Dashboard

In the upper part (1), it is possible to select the week and the datasource which, in our case, refers to voice calls. In addition to these two, it is also possible to select the k-shell level (2), which by default always corresponds to the maximum level of the network associated with the selected week. Due to the high amount of data, the networks developed consist of thousands of nodes and millions of connections. By using this technique it is possible to filter out a large portion of the nodes and select only part of the most important ones. If the maximum level is not sufficient, the user can select a lower k-shell level. This way, it is possible to adapt the visualization to the computational resources at hand.

In the networks shown in the dashboard (6), each circle represents a client, and a connection between two clients indicates that at least one call was made between them. In these connections, it is possible to analyse the direction of the call and the number of times it occurred during the selected period. On the other hand, the radius of the node is proportional to the level of influence of each client, meaning that the larger the radius, the greater the influence.

If the user clicks on one of the nodes displayed in the network, the sub-network associated with that client will be displayed. A customer subnet consists of all the nodes that are within a radius of 2 from the selected node. These subnets allow us to analyse the clients that are directly related to the selected node and how they relate to each other. The Figure 4.2 shows the sub-network associated with customer 190976 where 3 influencers are identified. The node in yellow represents the node selected by the user. In this representation it is also possible to observe a large number of nodes with a higher level of transparency. When this happens, it means that these nodes only made or received calls, being their

betweenness value zero. This means that they are not considered for the calculation of influencers as already discussed.

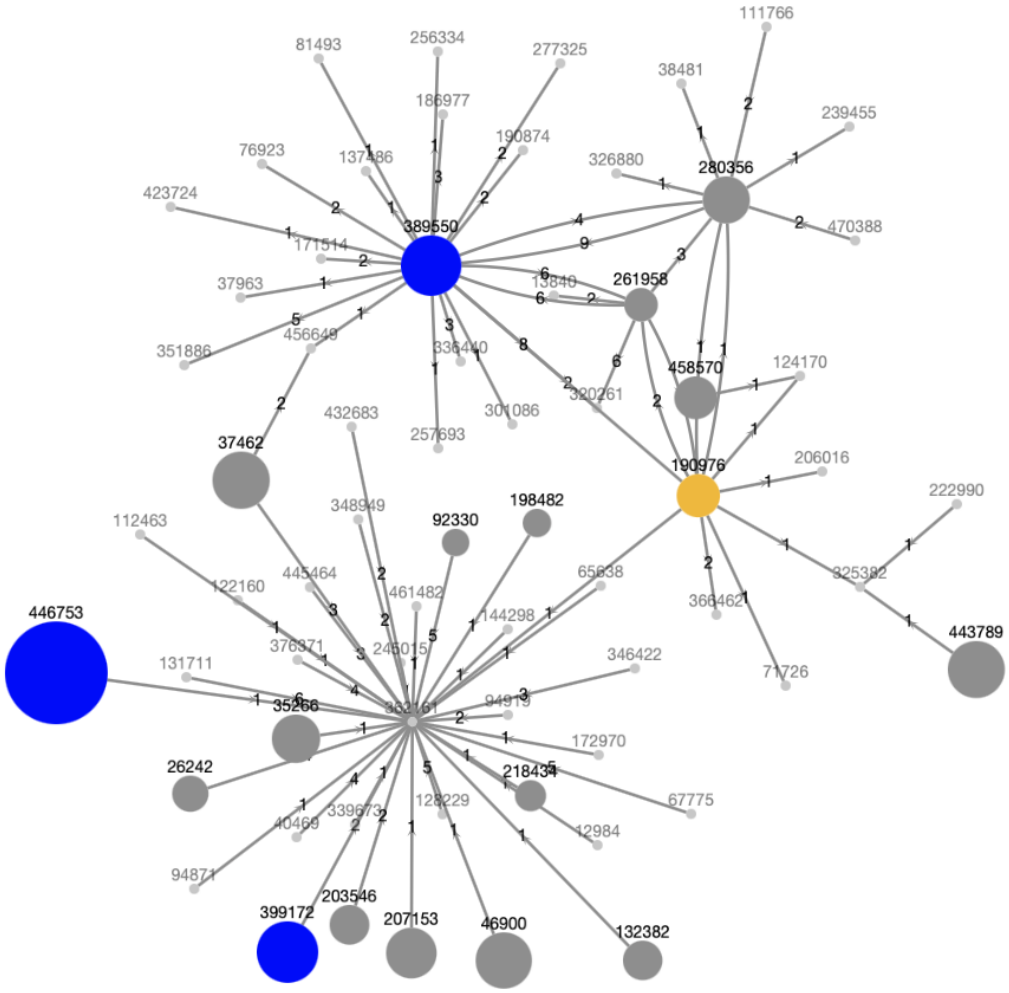

Figure 4.2: Client 190976 - Partial Network with Top 3 Influencers

In the box on the right hand side (3), it is possible to select what we want to analyse: The network influencers or the clients mobile operator. If we select the first option, the influencers are computed and represented with the blue colour. By default, this is obtained by using the expression 4.1 shown above, however, the metrics used in the calculation can be individually selected (4). This allows us to analyse the influence that each metric has on the final result. In addition, it is also possible to select whether the calculation is done on a local or global scale (5), i.e., considering the entire network or only the subnet being displayed at the time. By default, the global network is considered.

If the mobile operator option is selected, each node will be painted in the colour that represents its mobile operator. In order to make the representation clearer, five colours were used:

- Blue - MEO;
- Orange - NOS;
- Red - Vodafone;

- Green - Others;
- Grey - Undefined;

Figure 4.3 shows an example for the customer 71639. In this representation, the radius of the nodes still represents the level of influence, but the influencers are not identified. In this case, the colours represent the customer's mobile operator according to the colours that were defined above. The node in yellow represents the node that was selected by the user.

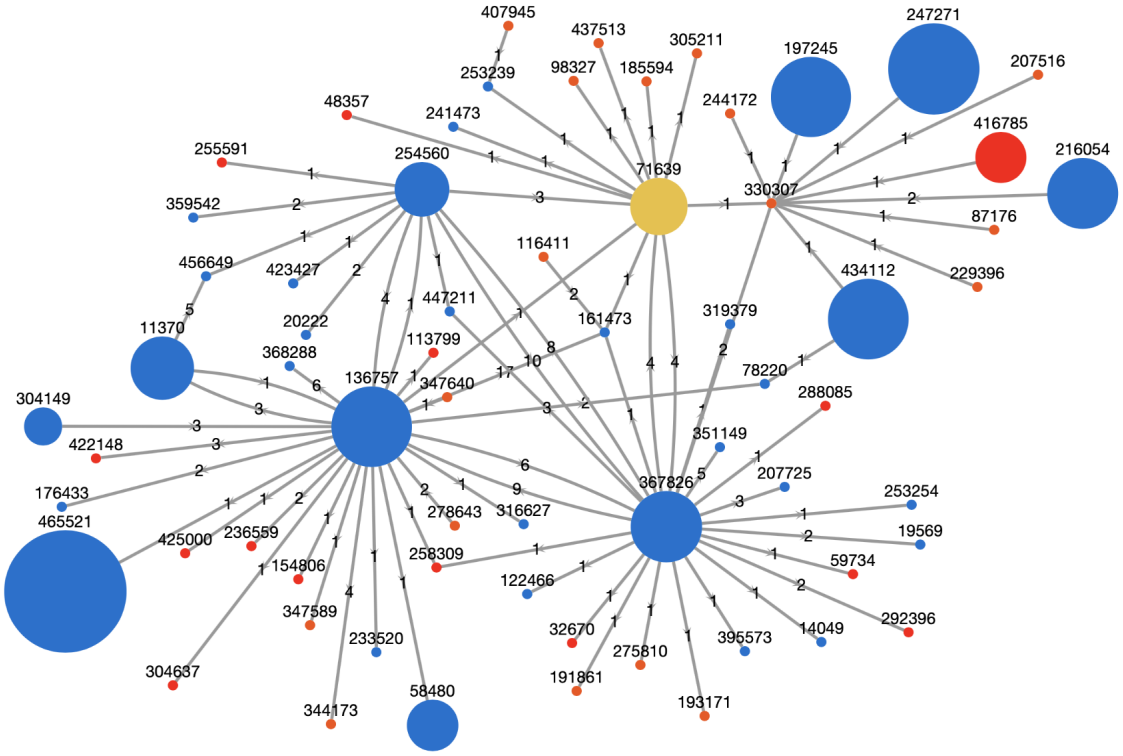

Figure 4.3: Client 71639 - Partial Network with Mobile Operators

4.4 Database

To store all the operations described in the previous sections, two tables were created in the database: The Networks table and the Client Scores table. The structure of these tables is shown in Figure 4.4. The developed dashboard is supported by the Networks table. This table stores all information about the networks created as well as the results of the metrics used to calculate influencers in JSON format. We decided to store everything in this data type because of the way the results are returned by the methods used. Since the results come in dictionary form, converting them to JSON is done straightforwardly for easy reading and writing. Also, to ensure performance when using the dashboard, the results of the sums of all combinations of the metrics available are also stored in this table. This is important since in the dashboard developed it is possible to select the metrics the user wants to use for the calculation of influencers. By having all the combinations of metrics already calculated, we ensure that the results are obtained in real time. In this table are also stored the classification results per mobile operator of all numbers.

The client scores table contains the information itself which can be used by the telecommunications company. Each record in this table is associated with a client and a particular week or time interval and contains the scores obtained for each metric in that week. Besides the metrics, it also contains the K-shell group to which that customer belongs, the maximum K-shell level of that particular week and the corresponding mobile operator.

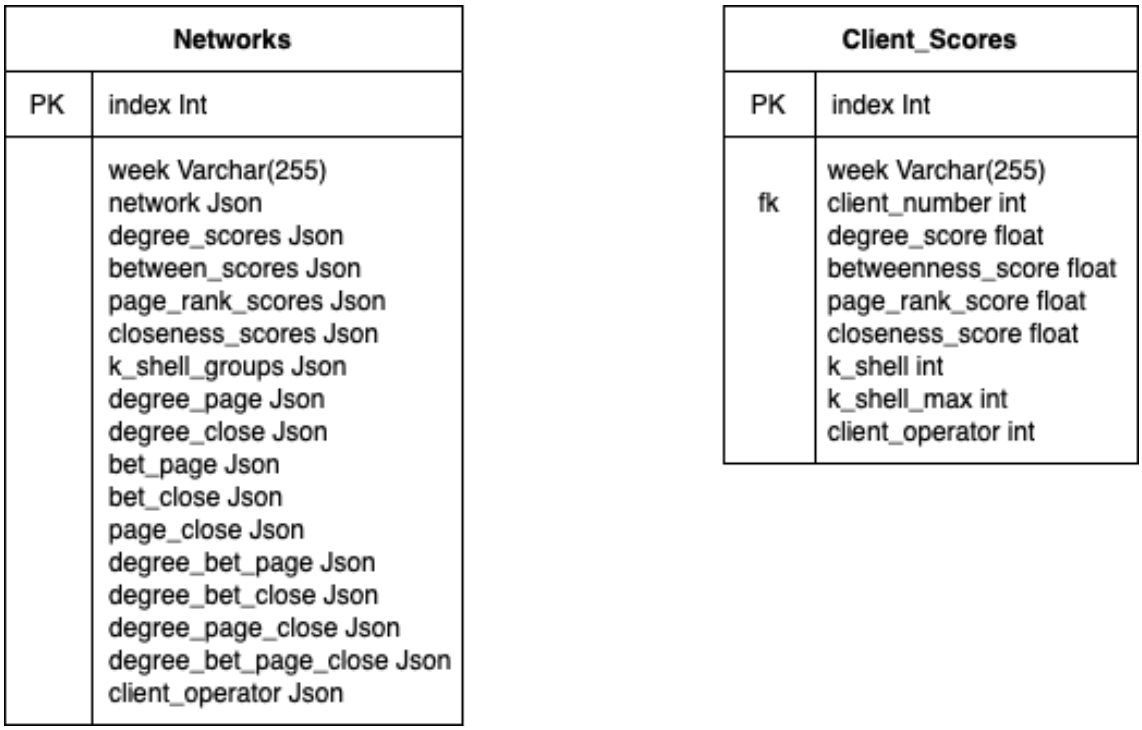

Figure 4.4: Influencers - Diagram of the Created Tables

It is important to mention that this is not the final structure of these tables. As we wanted to store the results of the detection of the relationship groups also in these tables, small changes had to be made. These transformations and the final diagram of the tables created are presented in section 5.5.

4.5 Discussion

Using the model developed we were able to study telephone records and through the interactions that this data reveals, it was possible to segment customers by level of influence. This influence is calculated using different metrics that analyse different network characteristics.

Throughout the development of the model, some difficulties were encountered, especially in the calculation of some metrics. Despite this, we managed to find alternatives that allowed us to lower their complexity in order to improve the performance of the whole model.

The results obtained are very useful for telecom companies because it enables them to obtain certain competitive advantages over other companies operating in this area. If this classification is correct and we are effectively identifying the most influential individuals, the company knows that it has to treat these individuals more carefully, being able to offer them certain products or services in order to guarantee their satisfaction and a good opinion about the company. As network influencers, these individuals tend to relate to

many people, which can be useful from a new product diffusion point of view and buzz marketing.

Through the developed tool it was possible to validate the results in a visual way. This validation process showed that the results are in accordance with the nature of the metrics used. The results obtained were also partially validated with the company team which led us to draw the same conclusions. Due to the nature of the problem, it is difficult to find data that proves that a particular individual is influential in their environment. Nevertheless, diligences are being taken so that the evaluation can be carried out in the next few months through surveys or other tools. Through these efforts that are being conducted for the validation of this model, we can conclude that the elements of the company team consider that the results produced by the developed model are relevant enough to the point of wanting to validate them in the field.

Chapter 5

Social Groups Detection

The household detection model constitutes one of the main goals of this work. Through the use of community detection techniques, it is possible to split a network of interactions into smaller relationship groups. In our model, this division was made through the use of an algorithm widely used throughout the literature and which is efficient when applied to large networks. Despite this, the algorithm used presents a problem that can compromise the results produced. This chapter walks the reader through the components that comprise this model and explains the decisions behind their design. This chapter also describes the validation process performed and their results.

5.1 Classification Model

As in the model developed for influencer detection, the construction of a network of interactions is essential. However, unlike the latter, we decided to create a network consisting of the traffic collected during the two months that made up our dataset.

Although the model is independent of the defined time interval, by increasing the time window of the developed networks, we are ensuring that all connections between individuals are considered. For example, if two individuals, A and B are close to each other in real life but their mutual call log says that they only talk from the first week of October, they would end up not being placed in the same group as there would be no evidence of a relationship between them in the networks created during September. Although the weekly networks are useful from a developmental point of view, they end up raising this issues of not including important relationships between individuals. On the other hand, by using all the data, we are also ensuring that the model performs well and can handle a considerable amount of data.

The network was constructed using the same technique described in Section 4.1. The characteristics of this network are described in the following table:

Number of Nodes	420357
Number of Edges	973409
Average Node Degree	4
Average Edge Weight	20

Table 5.1: Two Month Network Characteristics

Through the analysis of the values presented in the Table 5.1, we would like to highlight the average degree of each node being 4 meaning that each customer on average interacts with four different people. Also regarding the average weight of each connection, the value presented informs us that during the period of two months, the average number of calls between two clients is 20, giving a weekly average of two and a half calls.

After the creation of the network it was necessary to investigate what would be the best technique to divide the network into groups. During the selection process we aimed to select an algorithm that would be efficient and prepared to support large networks. After reviewing the literature, we came to the conclusion that the Louvain algorithm would be the best option due to its performance in large-scale networks [19, 38, 51, 52].

Despite its efficiency, the Louvain algorithm, being based on modularity optimization, has some issues when computing smaller communities. Networks with high modularity have dense connections between the nodes within communities and sparse connections between them. However, Fortunato and Barthelemy [53] showed that modularity suffers a resolution limit, failing to identify communities smaller than a scale that depends on factors such as the size of the network and the degree of interconnectedness of the detected groups.

Since this factor could become decisive in the results obtained, it was necessary to adapt Louvain's algorithm to our problem. Thus, we implemented an algorithm that we call "Louvain Recursive Algorithm" where we try to divide large communities into smaller ones according to a given threshold to overcome the resolution limit associated with modularity. Figure 5.1 shows the flow chart of this algorithm.

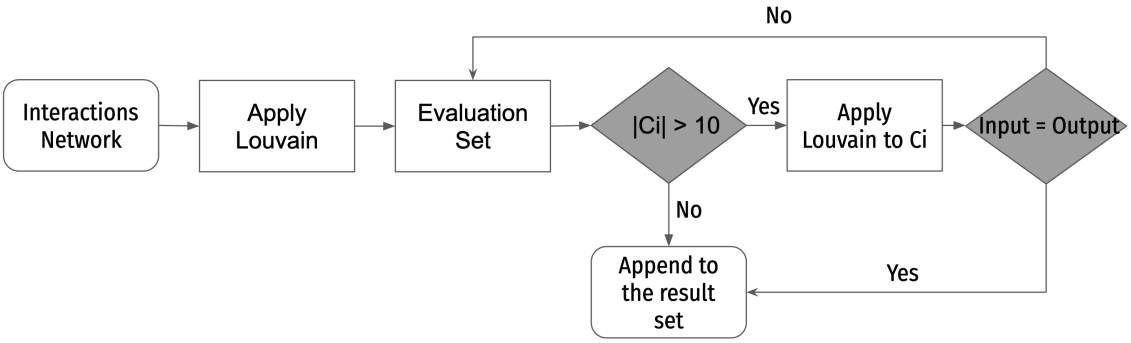

Figure 5.1: Recursive Louvain Algorithm

First we start by applying the Louvain Algorithm to the interaction network built earlier. The result of this step is a set of communities of variable size where each client number is associated with only one community. Then, the size of each of the resulting communities is evaluated: If the community size |Ci| is less than the threshold T, then that community respects the defined size and is stored in the result set. If the community size is greater than the threshold, we try to split the community into smaller ones, by applying the Louvain algorithm. At this stage, there may be communities that cannot be broken down into smaller communities, i.e., there is no gain in modularity value. Thus, if the resulting community is equal to the input community, it is added to the result set. Otherwise, the resulting communities are added to the evaluation set and the whole process is repeated. To complement this explanation we also present the pseudo-code of the algorithm:

Algorithm 1 Recursive Louvain Algorithm

```
G ← InteractionsNetwork
T ← SizeThreshold
EvaluationSet ← Louvain(G)
while EvaluationSet is not empty do
    for C_i in EvaluationSet do
        if | C_i | > T then
            Output ← Louvain(C_i)
            if Output is equal to C_i then
                ResultSet ← ResultSet + Output
            else
                EvaluationSet ← EvaluationSet + Output
            end if
        end if
    end for
end while
```

5.2 Results

The algorithm presented in the previous section was applied to the two-month network and the results obtained can be found in the following table:

Total number of communities	31389
Total number of communities smaller than 5 elements	4983
Total number of communities with size greater than 5 and less than 10 elements	9640
Total number of communities with size greater than 10 and less than 20 elements	11736
Total number of communities with size greater than 20 elements	5030
Average size	13
Size of the biggest/smallest community	616/1

Table 5.2: Recursive Louvain Algorithm Results

As shown, the algorithm produced 31389 communities for the two months of traffic. Analyzing the results in a more detailed way and taking into account the objectives of the decomposition of the network into communities, we can highlight two fields from those presented above. First, we consider that the number of communities with less than 5 elements can be representative of households. Taking into account that the number of elements per household in our country is on average 2.5 individuals [54], we consider this to be an acceptable number. As Coimbra is known as the city of students, we also highlight the average size of the communities. Usually, young people tend to relate to a lot of people being work or university colleagues. In this way, we consider that communities with a size around this value, may very well translate relationship groups, i.e., work or social ones.

When analyzing the set of results, we found the existence of two communities with a high number of elements, the largest consisting of 616 elements. By analyzing the constituent

elements, we found that the numbers that received a large volume of calls during the period under study were inserted in these communities. This type of outliers are produced since these numbers are connected to a large number of customers in the interaction networks. These individuals may well be couriers for companies such as Glovo or Uber Eats or even taxi drivers, which explains why they receive a high number of calls.

5.3 Validation of Results

Although the results obtained were promising, they need to be validated. Thus, it was necessary to find a way to confirm the results with concrete data about the customers that make up our dataset. The process of validating results allows us to ensure that the model we are developing is producing results according to its purpose. Depending on the problem at hand, various approaches can be taken.

Due to the nature of our work, there are some challenges concerning validation, mainly due to strict data protection policies. Although the company that provided us with the data initially possessed information that facilitated this validation process, it could not be made available for research purposes. Thus, we had to look for other valid solutions.

One hypothesis that emerged, and that in the end ended up being the one used, was the application of data resulting from another project established between the University of Coimbra and the company associated with this work. This project main objective was the prediction of the home address of the company's customers based on data collected through their mobile phone usage (Voice Calls, Text Messages, GPRS usage, among others). Using clustering techniques, it was possible to use the locations of the towers associated with each record to predict an approximate location.

The file provided contains the addresses of 4627 customers. Although it contains data on a small portion of the total number of customers, these results were produced using the same dataset as the one used in this project, ensuring that the anonymised codes were the same.

We know that, people belonging to the same household share the same address. Taking this into account, our idea was to use the information contained in this dataset to validate the communities produced and infer the elements belonging to the same community that are actually close to each other. In order to better understand this validation process, the diagram in Figure 5.2 is presented.

Firstly, we begin by loading the validation data where each entry contains the client number and location. For each of these clients that is represented in this scheme by **Ci**, it is checked if it has an associated community since as only data regarding voice traffic was used, there may be numbers that ended up not being considered by our model. (For instance, numbers that only have text messages associated).

If **Ci** has an associated community, we fetch the elements belonging to that community. For each element **Cn** of that community we verify if there is validation data for this number, i.e., if the number **Cn** is in the validation dataset. If not, it cannot be validated and the algorithm moves on to the next element. If there is validation data, it is calculated the distance between client **Ci** and **Cn** using the *Haversine Distance*. This formula allows us to calculate the distance between two points on the surface of a sphere from the latitude

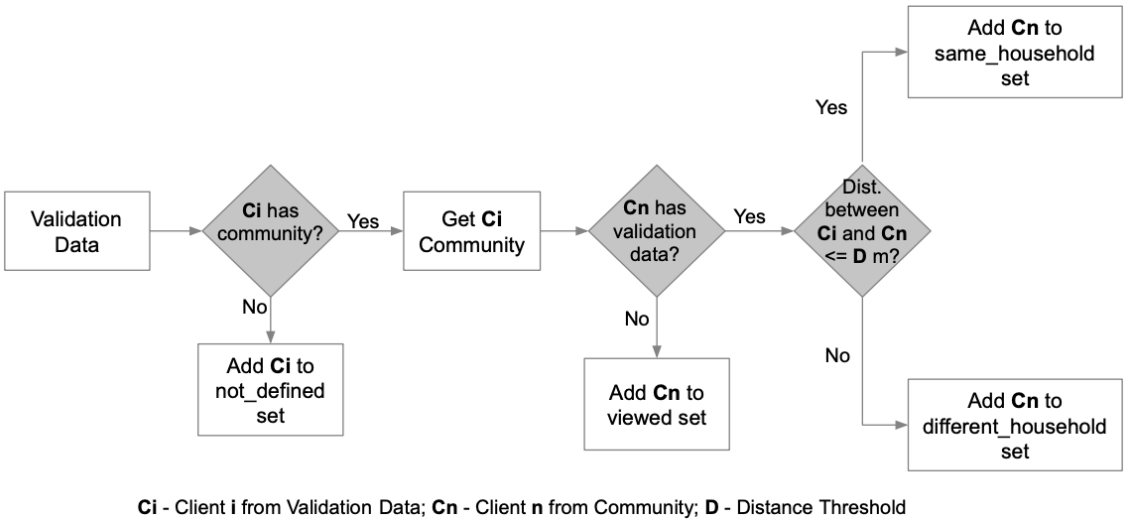

Figure 5.2: Validation Procedure

and longitude values [55]. The formula of *Haversine Distance* is shown below:

$$d = 2 * r \sin^{-1}\left(\sqrt{\sin^2\left(\frac{\Phi_2 - \Phi_1}{2}\right) + \cos(\Phi_1)\cos(\Phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \tag{5.1}$$

where r is the radius of the earth, d is the distance between two poitns, (Φ_1, Φ_2) is the latitude of the two points and (λ_1, λ_1) is the longitude of the two points respectively.

If the resulting distance is less than or equal to a threshold initially defined, we consider that both are close to each other and may belong to the same group. However, if the distance is greater, we consider that although they are classified in the same community, they should be in distinct ones.

For the experimental setup we decided to use two distinct values for the Threshold of distance (100 and 400 meters). With these values, we intend to analyse the influence of this distance on the results produced. The results obtained are presented in table 5.3.

Distance (m)	100	400
Number of clients	4627	
Number of clients with undefined community	259	
Number of clients with defined community	4368	
Number of distinct communities used in this validation	4261	
Communities where at least 2 elements are close to each other and classified in the same community	51	54
Communities where at least 2 elements are distant to each other and classified in the same community	52	49
Communities with at least 3 elements where at least 1 element is distant and 1 element is close and they are all classified in the same community	1	1

Table 5.3: Recursive Louvain Algorithm - Validation Results

For the purpose of analysing the results, we will only consider the values obtained for the 100m threshold since, being the lowest value, it allows a better accuracy of the distances between clients.

As can be seen in Table 5.3, the amount of data for validation is small considering the number of customers in our dataset, but it is already possible to draw some conclusions from the results obtained. In this validation process, 4261 distinct communities were used. Considering that we have 259 clients without a defined community, and considering the total number of clients with data for validation, we conclude that almost all clients are distributed in distinct communities.

Despite these results, we were able to obtain a small subset of clients where it was possible to draw some conclusions. Through this validation process, we were able to identify 51 communities where there are at least 2 elements with a close address to each other. Assuming that the addresses provided are correct, those elements classified in the same community that are close to each other may well represent one or part of a household. In opposition, 52 communities were found where our model grouped at least two clients in the same community and which in reality, they are distant between themselves. Finally there is a community where the two cases described occur. In this community we have at least three elements where one of them is close and the other one is distant, but they are all classified in the same community.

Given the size of the validation dataset, the obtained results should be interpreted with a grain of salt. However, they show us that the model we developed is working, even if further adjustments might be required. We believe that through other more complete sources of validation we would be able to draw more detailed conclusions from the results obtained.

5.4 Visualisation

To facilitate the evaluation of the communities produced, we included in the dashboard described in Section 4.3 the possibility to visualise the communities.

The selection can be made by customer phone number or community size. If the user chooses the first option, only the community belonging to the selected number will be displayed. Figure 5.3 illustrates a preview of the dashboard with the community associated with the number 61235. As with the Influencer Detection section, the radius of each node is proportional to the level of influence.

If the user chooses the second option, all communities of the chosen size will be displayed. This view only allows us to have the perception of the total number of communities with the defined size. Figure 5.4 shows all communities with size 4.

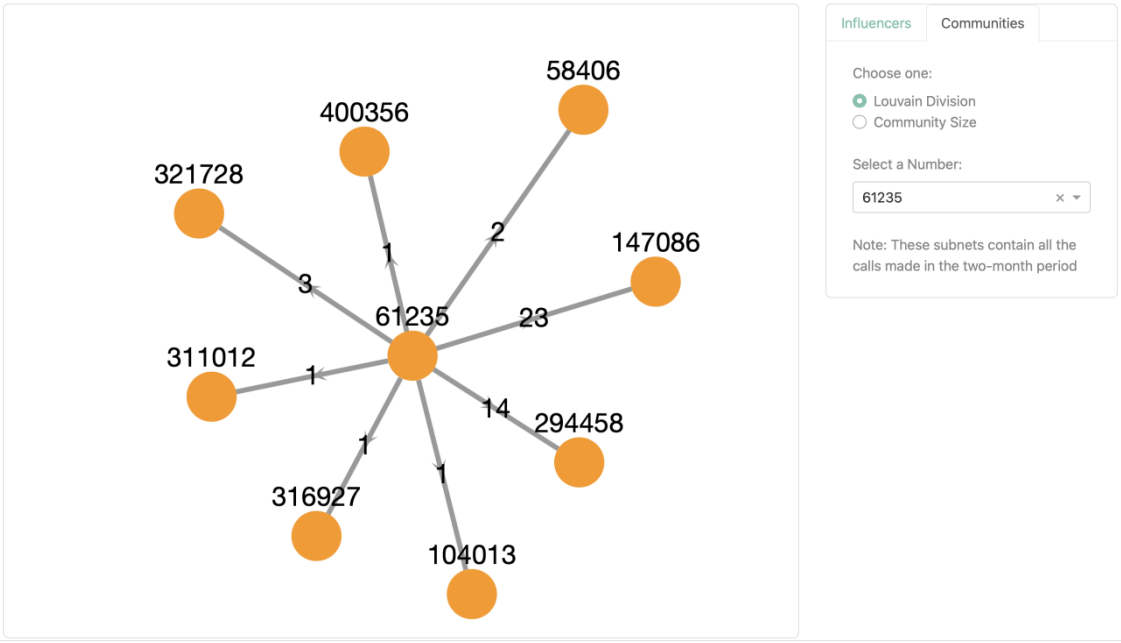

Figure 5.3: Dashboard Preview - Louvain Division Number 61235

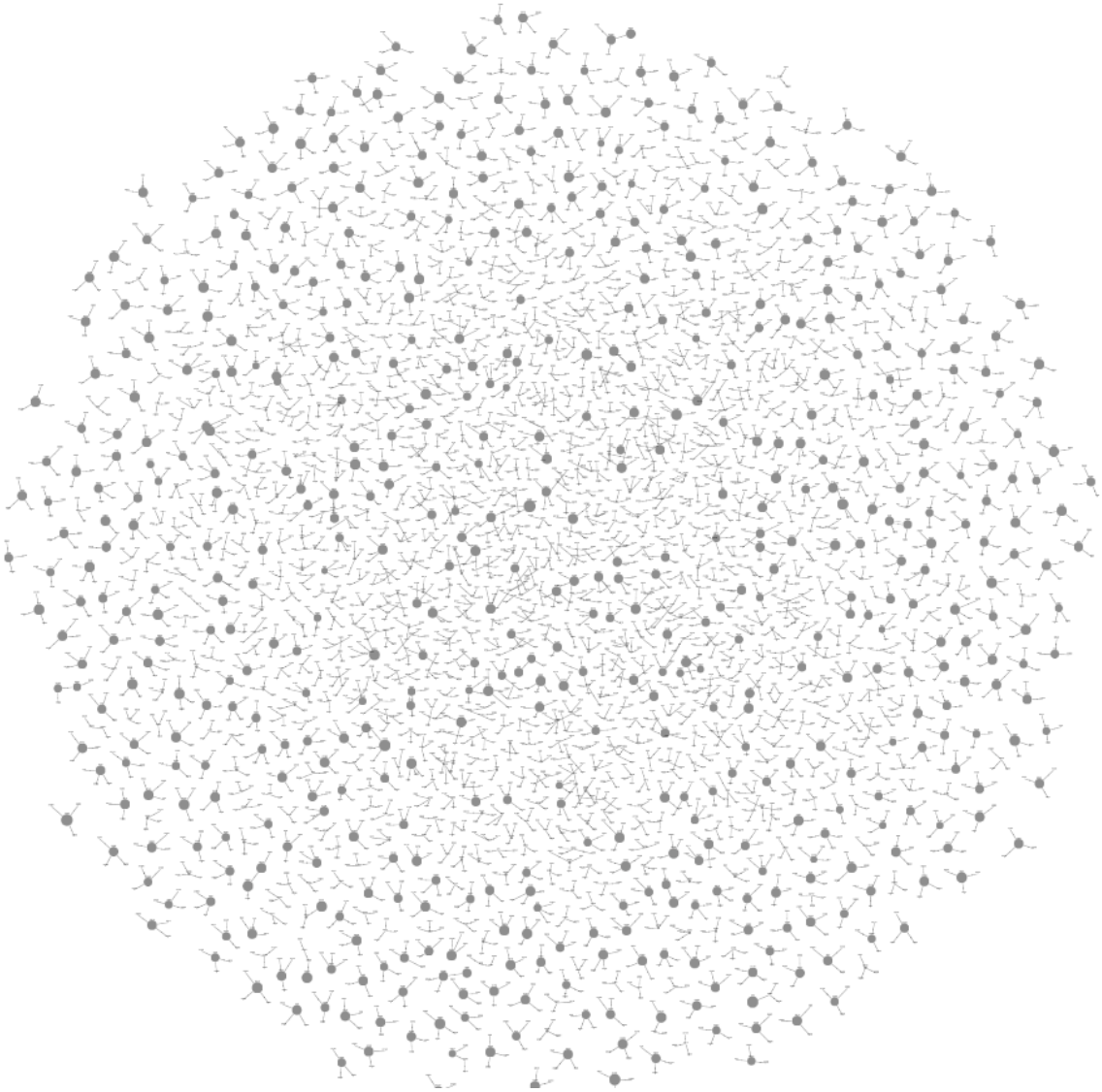

Figure 5.4: All Communities of size 4

5.5 Database

In order to store all the communities produced, a reference table was created where each entry comprises a community in JSON format, a unique identifier (integer index) and the time period used. This last field is useful because it allows communities produced using distinct time intervals to be stored in the same table.

In the Client Scores table referenced in the section 4.4, a column corresponding to the community was added. In addition to all the fields referred above, each record in this table has an identifier for a community to which the client belongs. By using this unique identifier, we simplify the representation and establish a link between the two tables.

The following figure shows the fields and data types of the tables created for the storage of all the data required for the operation of the models, storage of results and the running of the dashboard developed. The aim has always been to store data in a simple and structured way in order to simplify possible queries that can be done on the data.

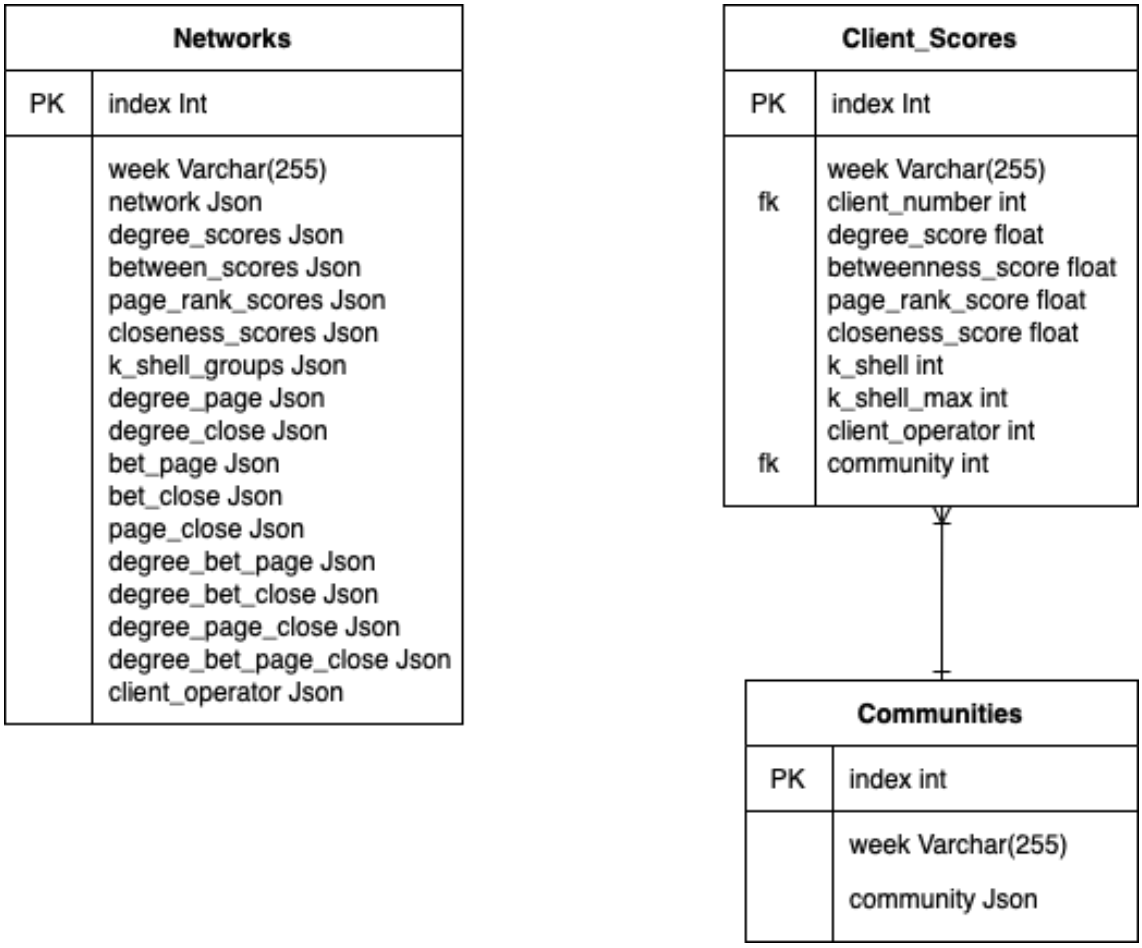

Figure 5.5: Final Diagram of the Created Tables

This page is intentionally left blank.

Chapter 6

Deployment and Integration

This work was conducted under an industry research project established between the CISUC and a telecommunications company. Since the beginning of the project's development, we were told that this company aimed to deploy and use the project's results in their environment. Taking into account the market in which they operate, by putting the developed models into production, they would be able to obtain the information and advantages described in the previous sections .

On the company's side a team was allocated, which despite being in charge of other projects, had the objective of following the whole development process of this work. For this reason, fortnightly meetings were held where all the advances were presented and the different approaches taken were discussed. In this way, it was possible for them to actively participate in the decisions made and to become familiar from an early stage with the techniques used.

As this work progressed, several technical reports were developed containing the full description of the models developed. In addition, it also had the descriptions of the datasets used and all the operations performed on the data, as explained in this document. The main objective of these reports was to document the whole process and the architecture of the models so that the understanding of the source code provided could be more efficient.

The delivery of the source code was done in two stages. In the first, the model for influencer detection and mobile operator identification was made available. In the second delivery, the model regarding relationship groups was provided. After these deliveries, there were some meetings to visit and analyse the source code made available. The main goal of these meetings was to explain how the code is structured and to clarify any doubts that may have arisen from the company's team. At this stage, the team allocated to this project intended to put the models into operation in their environment, to evaluate the best way to put them into production.

Performance Issues and Database

In general, the developed models have a good performance being prepared to receive large amounts of data. However, during the code visit meetings, some problems in terms of performance and incompatibilities were discovered that had to be solved.

Betweenness is the computationally heaviest metric that makes up the influencer ranking model. Through weekly divisions, we were able to calculate this metric without any problems, but when applied to the two-month network created to calculate communities, we found that our implementation was not efficient enough. To overcome this problem, we

decided to explore a GPU implementation. Our implementation for the two months takes about 12 days running on CPU while on GPU, for the same network, it takes about 50s resulting in a very high performance gain.

Rapids' implementation of CuGraph [56] was used since it is compatible with NetworkX graphs. This library does not allow direct graphs to be used, so a conversion of the network to undirected had to be made. This transformation makes the results produced a little different from those we had obtained with the CPU implementation. Therefore, a Python script was developed evaluate the difference of positions between the two sets of results (The results were ordered by Betweenness value). After performing this analysis we concluded that the differences were not significant and that the gain in performance justified it (These results can be found in Appendix B). Thus, in the final version of the source code, the GPU version of the Betweenness was provided which can be beneficial in terms of performance on a large scale.

Another problem found in the provided code was related to incompatibilities in database operations. As already mentioned, we decided to use a PostgreSQL database, however, the company in its environment only uses Oracle databases. During the implementation of the code on the company's system, some incompatibilities were found in the queries for searching and storing data, since, despite both being relational databases, the way they operate is not the same. As it was not possible to solve the problems in the code visit meetings and as we could not have access to the company's environment, we decided to replicate our database in Oracle and change all the operations that are made to be compatible with this type of database. This process was somewhat lengthy since it was necessary to install all the necessary software and learn some basics about Oracle database creation and handling. Despite the difficulties encountered, we managed to find and solve problems related to incompatibilities between the two types of databases having everything working correctly.

Chapter 7

Conclusion

Companies, especially those operating in the telecommunications area, are under more and more pressure to remain competitive, seeking more and more solutions that allow them to obtain advantages. In recent years, they have invested a large amount of resources and effort in customer relationship management through segmentation techniques to tailor their products to customer needs.

In this work we explored the possibilities of using Call Detail Records with respect to customer segmentation by telecommunication companies. Specifically, two models were developed that allow detecting relationship groups and measuring the level of customer influence based on these sources. Furthermore, a way to classify these customers based on their mobile operator was presented, which can be very useful from a marketing strategy design point of view. By creating these models, we aimed to improve the knowledge that a telecommunications company has about its customers not only to increase its productivity and competitiveness levels but also to improve the way it approaches its customers and the relationship it has with them.

The influencer detection model allowed us to identify the most influential individuals in a network of interactions created through call logs. This classification was based on 4 state of the art metrics that evaluate different network characteristics and associate an influence score to each element of these networks. Through a visual tool purposely developed for this work we were able to analyse and validate the results obtained to confirm that the results are in accordance with what was expected. Due to the nature of the problem, it is difficult to find real data which specify that a particular individual is influential. Thus, the process of validating this model was done solely on the basis of our perception of the structure of the networks and throught feedback of domain experts. Despite this, efforts are being made by a company to conduct a validation process in the upcoming months. They intend through surveys and other tools in the field, to confirm that the results produced are reliable which underlines the importance of this type of model.

Based on an adaptation of an algorithm from literature, the second model developed was aimed at detecting relationship groups, whether family or social ones. The results were validated using the addresses of some of the customers who were part of our interaction network. Despite the limitations found in this validation process, the results showed that the model developed is working and it is possible to detect households, supporting our hypothesis that individuals belonging to the same household establish a proximity relationship that can be confirmed through their mobile traffic.

Our work constitutes the first step in the use of call detail records for customer segmen-

tation. Companies previously collected this type of data only to extract technical details and, through these models, we were able to develop methods to extract even more value from these sources.

In conclusion, we consider that the developed models were able to perform their tasks effectively and the results support our initial hypothesis.

7.1 Future Work

When analysing the advances achieved throughout the development of this work, we believe that new opportunities for research were created.

The data used in this work to create the interaction networks relate only to call records. We believe that the introduction of other sources, such as SMS traffic, would allow the addition of a new layer of interactions between individuals, potentially producing a more detailed analysis of their relationships.

The algorithm used in this work was chosen due to the good results that it has obtained in many research projects and due to its overall performance. We consider that it would be interesting to apply different community detection algorithms and establish a comparative analysis of the results obtained.

Finally, there are many analyses that can be done on the detected groups. It would be interesting to gather other data sources to segment the groups found in order to assess, for example, demographic characteristics of the individuals that constitute them. We believe that this type of segmentation may be important to complement the information that companies have about their customers and adapt the way they approach them. Through this segmentation, it would be easier for companies to classify the groups detected as social, work or family groups.

References

[1] Nour Al-molhem, Yasser Rahal, and Mustapha Dakkak. Social network analysis in telecom data. *Journal of Big Data*, 6, 11 2019. doi: 10.1186/s40537-019-0264-6.

[2] Greg Metz Thomas Jr. Building the buzz in the hive mind. *Journal of Consumer Behaviour*, 4(1):64–72, 2004. doi: https://doi.org/10.1002/cb.158. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/cb.158`.

[3] B. Csáji, A. Browet, V. Traag, J. Delvenne, E. Huens, P. Van Dooren, Z. Smoreda, and V. Blondel. Exploring the mobility of mobile phone users. In *Physica A: Statistical Mechanics and its Applications*, 2012.

[4] Ehsan Jokar and Mohammad Mosleh. Community detection in social networks based on improved label propagation algorithm and balanced link density. *Physics Letters A*, 383(8):718–727, 2019. ISSN 0375-9601. doi: https://doi.org/10.1016/j.physleta.2018.11.033. URL `https://www.sciencedirect.com/science/article/pii/S0375960118311794`.

[5] Siti Nurulain Mohd Rum, Razali Yaakob, and Lilly Affendey. Detecting influencers in social media using social network analysis (sna). *International Journal of Engineering & Technology*, 7:950, 12 2018. doi: 10.14419/ijet.v7i4.38.27615.

[6] Mehdi Azaouzi, Delel Rhouma, and Lotfi Romdhane. Community detection in large-scale social networks: state-of-the-art and future directions. *Social Network Analysis and Mining*, 9, 05 2019. doi: 10.1007/s13278-019-0566-x.

[7] Colin Shearer. The crisp-dm model: The new blueprint for data mining. *Journal of Data Warehousing*, 5:13–22, 10 2000.

[8] Gregory Piatetsky. Crisp-dm, still the top methodology for analytics, data mining, or data science projects. `https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html`, 2014. Accessed 2021-09-27.

[9] Shazia Tabassum, Fabiola S. F. Pereira, Sofia Fernandes, and João Gama. Social network analysis: An overview. *WIREs Data Mining and Knowledge Discovery*, 8(5):e1256, 2018. doi: https://doi.org/10.1002/widm.1256. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1256`.

[10] Maksim Kitsak, Lazaros Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H. Stanley, and Hernan Makse. Identification of influential spreaders in complex networks. *Nature Physics*, 6, 01 2010. doi: 10.1038/nphys1746.

[11] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. URL `http://ilpubs.stanford.edu:8090/422/`. Previous number = SIDL-WP-1999-0120.

[12] David Easley and Jon Kleinberg. *Networks, Crowds, and Markets: Reasoning About A Highly Connected World*. 07 2010. ISBN 978-0-521-19533-1. doi: 10.1017/CBO9780511761942.

[13] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002. ISSN 0027-8424. doi: 10.1073/pnas.122653799. URL `https://www.pnas.org/content/99/12/7821`.

[14] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Phys. Rev. E*, 70:066111, Dec 2004. doi: 10.1103/PhysRevE.70.066111. URL `https://link.aps.org/doi/10.1103/PhysRevE.70.066111`.

[15] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E*, 76:036106, Sep 2007. doi: 10.1103/PhysRevE.76.036106. URL `https://link.aps.org/doi/10.1103/PhysRevE.76.036106`.

[16] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74:036104, Sep 2006. doi: 10.1103/PhysRevE.74.036104. URL `https://link.aps.org/doi/10.1103/PhysRevE.74.036104`.

[17] J. Xie and B. K. Szymanski. Community detection using a neighborhood strength driven label propagation algorithm. In *2011 IEEE Network Science Workshop*, pages 188–195, 2011. doi: 10.1109/NSW.2011.6004645.

[18] Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. In pInar Yolum, Tunga Güngör, Fikret Gürgen, and Can Özturan, editors, *Computer and Information Sciences - ISCIS 2005*, pages 284–293, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-32085-2.

[19] Vincent Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics Theory and Experiment*, 2008, 04 2008. doi: 10.1088/1742-5468/2008/10/P10008.

[20] P. Santhi Thilagam. *Applications of Social Network Analysis*, pages 637–649. Springer US, Boston, MA, 2010. ISBN 978-1-4419-7142-5. doi: 10.1007/978-1-4419-7142-5_29. URL `https://doi.org/10.1007/978-1-4419-7142-5_29`.

[21] Maryam Fatemi and Laurissa Tokarchuk. A community based social recommender system for individuals & groups. 09 2013. doi: 10.1109/SocialCom.2013.55.

[22] Marcel Salathé and James Jones. Jones, j.h.: Dynamics and control of diseases in networks with community structure. plos comput. biol. 6(8), e1000736. *PLoS computational biology*, 6:e1000736, 04 2010. doi: 10.1371/journal.pcbi.1000736.

[23] Wei Chen, Yajun Wang, and Siyu Yang. Efficient influence maximization in social networks. volume 199-208, pages 199–208, 01 2009. doi: 10.1145/1557019.1557047.

[24] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 137-146, 07 2003. doi: 10.1145/956750.956769.

[25] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne Vanbriesen, and Natalie Glance. Cost-effective outbreak detection in networks. volume 420-429, pages 420–429, 01 2007. doi: 10.1145/1281192.1281239.

[26] Greg Satell. How the nsa uses social network analysis to map terrorist networks. `https://digitaltonto.com/2013/how-the-nsa-uses-social-network-analysis-to-map-terrorist-networks/`, 2013. Accessed 2021-01-07.

[27] Amrani Amine. Pagerank algorithm, fully explained. `https://towardsdatascience.com/pagerank-algorithm-fully-explained-dc794184b4af/`, 2020. Accessed: 2021-01-07.

[28] R. Ahas, S. Silm, O. Järv, E. Saluveer, and M. Tiru. Using mobile positioning data to model locations meaningful to users of mobile phones. In *Journal of Urban Technology, 17: 1*, page 3 — 27, 2010.

[29] Sibren Isaacman, Richard Becker, Ramón Cáceres, Stephen Kobourov, Margaret Martonosi, James Rowland, and Alexander Varshavsky. Identifying important places in people's lives from cellular network data. pages 133–151, 06 2011. ISBN 978-3-642-21725-8. doi: 10.1007/978-3-642-21726-5_9.

[30] L. Tongsinoot and V. Muangsin. Exploring home and work locations in a city from mobile phone data. In *2017 IEEE 19th International Conference on High Performance Computing and Communications; IEEE 15th International Conference on Smart City; IEEE 3rd International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pages 123–129, 2017. doi: 10.1109/HPCC-SmartCity-DSS.2017.16.

[31] Yan Leng. *Urban Computing using Call Detail Records: Mobility Pattern Mining, Next-location Prediction and Location Recommendation.* PhD thesis, MASSACHUSETTS INSTITUTE OF TECHNOLOGY, 2016.

[32] M. Vanhoof, F. Reis, and Z. Smoreda. Detecting home locations from CDR data: introducing spatial uncertainty to the state-of-the-art. 2018.

[33] KS Kung, K Greco, S Sobolevsky, and Ratti C. Exploring universal patterns in human home-work commuting from mobile phone data. 2014.

[34] Francesco Calabrese, Mi Diao, Giusy Di Lorenzo, Joseph Ferreira, and Carlo Ratti. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation Research Part C: Emerging Technologies*, 26:301 – 313, 2013. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2012.09.009. URL `http://www.sciencedirect.com/science/article/pii/S0968090X12001192`.

[35] Kevin Kung, Stanislav Sobolevsky, and Carlo Ratti. Exploring universal patterns in human home-work commuting from mobile phone data. *PloS one*, 9, 11 2013. doi: 10.1371/journal.pone.0096180.

[36] Thomas Louail, Maxime Lenormand, Oliva Garcia Cantu Ros, Miguel Picornell, Ricardo Herranz, Enrique Frias-Martinez, Jose Javier Ramasco, and Marc Barthelemy. From mobile phone data to the spatial structure of cities. *Scientific reports*, 4, 01 2014. doi: 10.1038/srep05276.

[37] Teodoro Dannemann, Boris Sotomayor-Gómez, and Horacio Samaniego. The time geography of segregation during working hours. *Royal Society Open Science*, 5:180749, 10 2018. doi: 10.1098/rsos.180749.

[38] Blaise Ngonmang, Emmanuel Viennet, and Maurice Tchuente. *Predicting Users Behaviours in Distributed Social Networks Using Community Analysis*, pages 119–138. 05 2014. ISBN 978-3-319-05911-2. doi: 10.1007/978-3-319-05912-9_6.

[39] Blaise Ngonmang, Emmanuel Viennet, and Maurice Tchuente. Local community identification in social networks. *Parallel Processing Letters*, 22, 03 2012. doi: 10.1142/S012962641240004X.

[40] Nathan Eagle, Alex Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences of the United States of America*, 106:15274–8, 09 2009. doi: 10.1073/pnas.0900282106.

[41] Xingguo Wu. Study on call detail records of family members based on classification model. 01 2016. doi: 10.2991/essaeme-16.2016.75.

[42] GDPR. GDPR - General Data Protection Regulation. `https://gdpr-info.eu/`. Accessed: 2021-10-15.

[43] Data Anonymization. Data Anonymization - Definition and Different Techniques. `https://bit.ly/2XIpwlp`. Accessed: 2021-10-20.

[44] Ashwin Machanavajjhala, Daniel Kifer, John Abowd, Johannes Gehrke, and Lars Vilhuber. Privacy: Theory meets practice on the map. In *2008 IEEE 24th International Conference on Data Engineering*, pages 277–286, 2008. doi: 10.1109/ICDE.2008.4497436.

[45] Synthetic Data. Synthetic Data - Key Benefits, Types, Generation Methods and Challenges. `https://bit.ly/3b8NxVO`. Accessed: 2021-10-22.

[46] Philip Kotler and Karen Fox. *Strategic Marketing for Educational Institutions*. 1998. ISBN 9788522411160.

[47] Khalid Rababah, Haslina Mohd, and Huda Ibrahim. A unified definition of CRM towards the successful adoption and implementation. 2010.

[48] PostgreSQL. PostgreSQL Documentation. `https://www.postgresql.org/docs/14/index.html`. Accessed: 2021-10-09.

[49] Plotly and Dash. Plotly and Dash Documentation. `https://dash.plotly.com`. Accessed: 2020-12-04.

[50] Networkx. Networkx Documentation. `https://networkx.org`. Accessed: 2021-05-17.

[51] Zhao Yang, René Algesheimer, and Claudio Tessone. A comparative analysis of community detection algorithms on artificial networks. *Scientific Reports*, 6, 08 2016. doi: 10.1038/srep30750.

[52] Andrea Lancichinetti and Santo Fortunato. Community detection algorithms: A comparative analysis. *Physical Review E*, 80(5), Nov 2009. ISSN 1550-2376. doi: 10.1103/physreve.80.056117. URL `http://dx.doi.org/10.1103/PhysRevE.80.056117`.

[53] Santo Fortunato and Marc Barthelemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences of the United States of America*, 104:36–41, 02 2007. doi: 10.1073/pnas.0605965104.

[54] PORDATA. PORDATA - Average Household Size in Portugal. `https://bit.ly/3j4qdNj`. Accessed: 2021-10-15.

[55] Mangesh Nichat. Landmark based shortest path detection by using A* Algorithm and Haversine Formula. 04 2013.

[56] Rapids - CuGraph. Rapids - CuGraph documentation. `https://docs.rapids.ai/api/cugraph/stable`. Accessed: 2021-10-15.

This page is intentionally left blank.

Appendices

Appendix A - Confidential

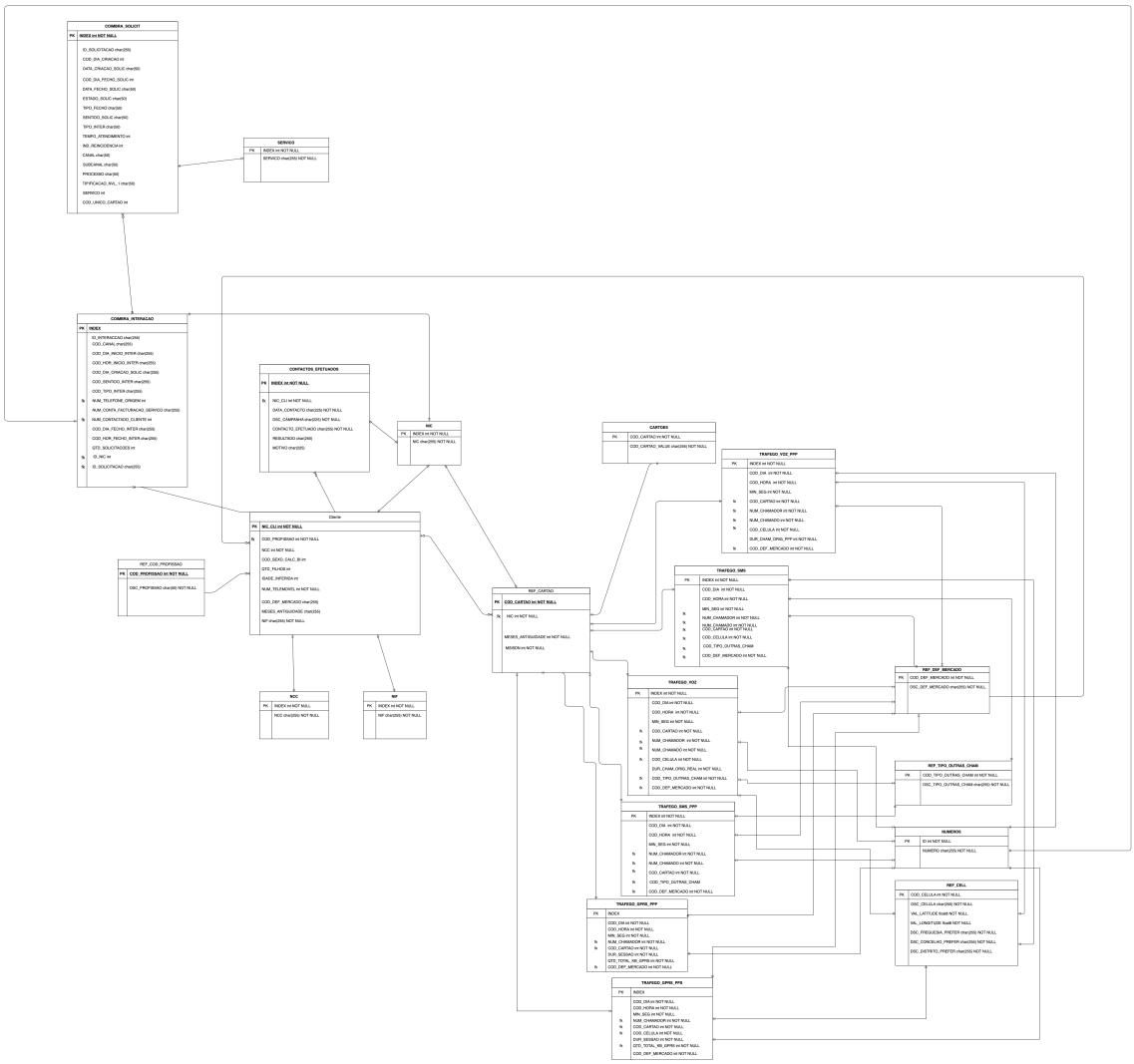

Figure 1: Data Model

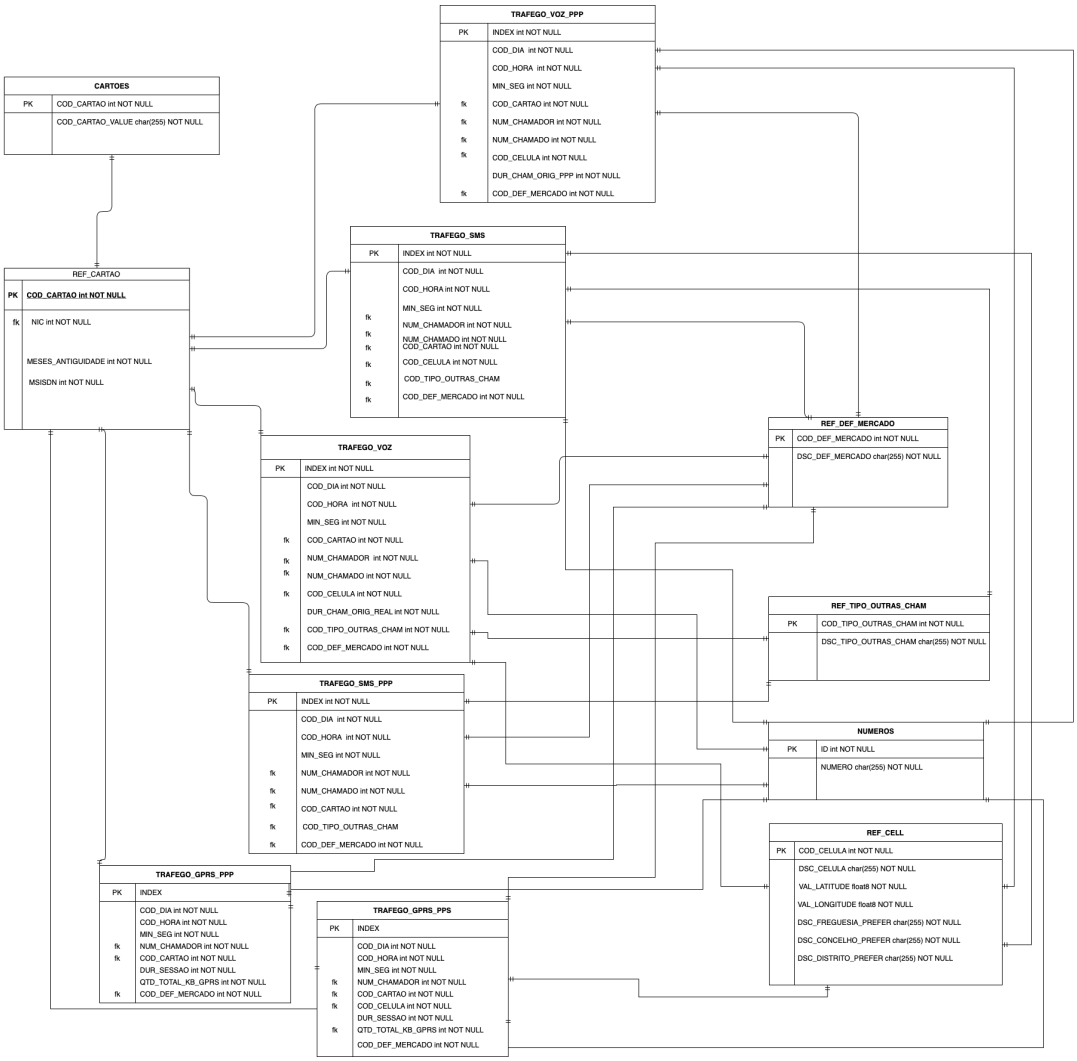

Figure 2: Segment of the Database Model

Appendix B

This graph shows an example of the analysis done on the betweenness values between Networkx and CuGraph methods for a set of 10 and 50 customers (Fig. 3 and Fig. 4, respectively). In this representation, the X-axis represents the customer number and the Y-axis the difference of positions in the ordered result of the two methods. As shown, the differences in positions are not significant.

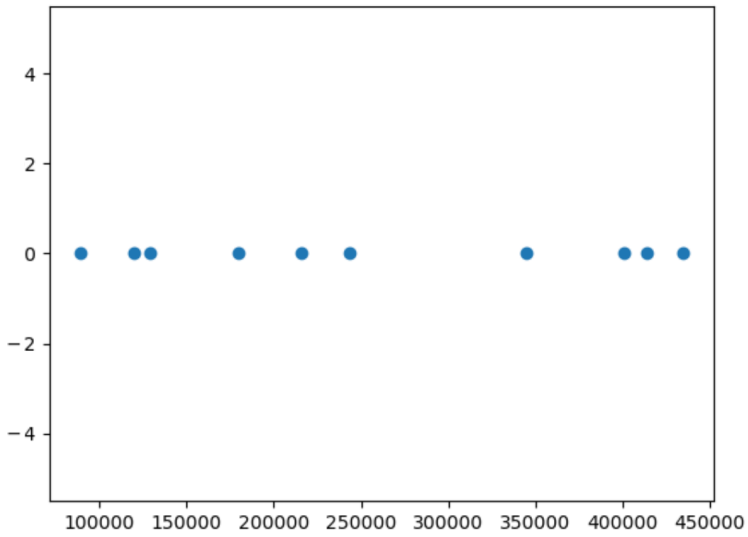

Figure 3: Betweenness Evaluation - 10 Customers

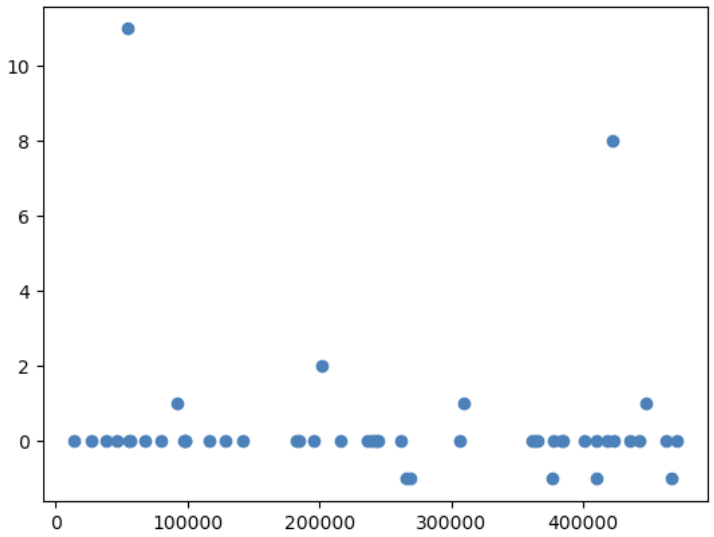

Figure 4: Betweenness Evaluation - 50 Customers