Learning About Clients From Call Graph

Beatriz Matos Júdice Gamito

Thesis to obtain the Master of Science Degree in

ELECTRICAL AND COMPUTER ENGINEERING

Supervisor: Prof. Cláudia Alexandra Magalhães Soares

Examination Committee

Chairperson: Prof. João Fernando Cardoso Silva Sequeira
Supervisor: Prof. Cláudia Alexandra Magalhães Soares
Members of the Committee: Dra. Sabina Zejnilovic

June 2018
Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.
Abstract

Based on a dataset provided by a telecommunications operator with fully anonymized information about mobile phone communications, the objective is to construct a mobile call graph and evaluate to which extent are network metrics informative for the prediction of age and gender of users from the call graph.

A mobile call graph is constructed to represent the social network of mobile phone users and facilitate the understanding of their communication behaviors. Extensive exploratory data analysis is performed to assess the quality of the data and investigate social strategies associated to the network structure. Since the state of the art is not reproducible, a preliminary approach with classical classification methods is proposed for gender and age inference in order to assess the added value of network metrics as features.

This thesis shows in detail that several social behaviors adopted by mobile phone users can be identified in call graphs, some of them in accordance to state of the art based on other datasets, thus proving the correlation between demographics and mobile phone communication behaviors within the network. Classification tree-based methods are tested with and without network features as predictors, with negligible performance difference. Therefore, even though the state of the art and the social strategies confirmed in this dataset show the merit of network structure in inference, the tree-based classifiers do not benefit much from it. This verification opens the door for exploring more complex graph-based models, such as probabilistic graphical models, which can better leverage the network structure for improved prediction.

Keywords

Classifiers, Communication behaviours, Demographic prediction, Mobile Call Graph, Social strategies, Probabilistic Graphical Models, Conditional Random Fields
1.1 Bias vs Variance tradeoff. Lower complexity models are prone to higher prediction errors, suffering from high bias. This means that the model failed to capture major features of the training data and so, does not allow for accurate predictions on other datasets. On the other hand, a higher complexity model tends to have high variance, which signifies an overfitting to the training set (therefore obtaining minor prediction errors), and consequent poor performance in a different data sample. A good model will have to assume a compromise between complexity and prediction error, as depicted in the graphic. source [1] 4

1.2 Example of a Bayesian Network (a) and a Markov Network (b). The graphical representation can be used to assess two different perspectives of the probabilistic graphical model. On the one hand, one can see the independencies induced from the graph: from (a), (B ⊥ C|A), (E ⊥ A|B,C), (D ⊥ C,E|B), (D ⊥ E|B), and from (b), (A ⊥ D|B,C), (B ⊥ C|A,D), where ⊥ stands for conditional independence.

On the other hand, the graph translates the factorization of the distribution such that
\[ P(A,B,C,D,E) = P(A|B)P(C|A)P(E|B,C)P(D|B) \]
in the case of the BN in (a) and
\[ P(A,B,C,D) = \frac{1}{Z} \phi_1(A,B)\phi_2(B,C)\phi_3(C,D)\phi_4(D,A) \]
in the case of the MN in (b). 5

1.3 Example of an ego network. A network is composed of vertices, the circles, and edges, i.e., the central node. Depicted as dark circles are the immediately connected vertices, the ego’s neighbors. An ego network further includes the connections among them. 7

1.4 Example of a factor graph. The \( X_i \) nodes are variable nodes, while the \( f \) are factor nodes. Connection between pairs of variable nodes are made through factor nodes, that represent the mathematical dependence of the variables. Factors can be related to a single variable, as is the case of \( f_1 \), and it is also possible to have more than one factor respective to the same pair of variables, as happens with \( X_1 \) and \( X_2 \) through \( f_2 \) and \( f_3 \). 7

1.5 Illustration of the graphical model proposed by Dong et al., where \( y_i \) is gender, \( z_i \) is age and \( x_i \) is the feature vector of user \( i \). Three different kinds of factors were defined: \( f(y_i, z_i, x_i) \) represents the relation between demographic variables and the set of features of user \( i \); \( g_l(y_i, z_i, y_j, z_j) \) represent the dyadic relation between users \( i \) and \( j \) and \( h_l(y_i, z_i, y_j, z_j, y_k, z_k) \) refers to the triadic relation among users \( i, j \) and \( k \). source [2] 11
2.1 Example of an undirected, unweighted graph. The path between two nodes is defined as the collection of edges, also referred to as hops, crossed when going from one node to the other. The shortest path between two nodes is the minimum numbers of hops necessary to go from one to the other. For example, there are two possible paths going from node A to node E: A-E and A-B-C-E. The shortest path between A and E has the size of 1 hop, whilst the shortest, but not unique, path from A to D has the size of 3 hops. Another relevant concept is the node degree – the number of edges connecting to one node. In this example, nodes A and B have degree 2, nodes C and E have degree 3 and nodes D and F have degree 1.

2.2 Degree distribution of the sample graph. The construction of the sample graph is detailed on Chapter 4. As seen in the literature, the degree for a mobile call graph typically assumes a heavy tailed distribution. When fitting a power law to this distribution the exponent obtained corresponds to 4.23, which is slightly smaller than the ones found in the literature for other undirected mobile call graphs.

4.1 Scheme of the format of a DS project. The DS project will be running on a program which will encompass different stages that go from importing the data, tidying it, understanding the data and communicating the findings. During the understanding stage, it is commonly necessary to transform, visualise and model the data iteratively, so that information can be extracted. Source: [3].

4.2 Distribution of the gender (a) and age (b) variables in the mobile call dataset. There is an overrepresentation of male over female and of older generations over younger generations of mobile phone users in the dataset, which will introduce an important bias factor in the data analysis.

4.3 Mean values of the variables number of calls (a) and number of sms (b) for each demographic group. It becomes evident that calling is the preferred communication method over exchanging sms and that females tend to communicate more often than men. Additionally, younger generations have a higher predisposition to use sms, even though they receive more communications than the ones they make, when compared to older generations, which prefer calling.

4.4 Average duration of a call calculated by gender and age. Even though the average duration of a call is well balanced throughout all demographic groups, there is a small tendency for older generations to participate in longer calls, when compared to younger ones.
4.5 Distribution of gender of a user’s contacts. Homophily has been confirmed in mobile call graphs, which means females would have more female contacts, while males would have more male contacts. Since the dataset is imbalanced, gender homophily can only be discerned for male contacts (b): the fact that there are mostly males in this network makes it so that women also have more male contacts. Even so, the percentage of female friends in considerably higher for female users (a), which could indicate the existence of general gender homophily in the mobile call graph.

4.6 Distribution of age of a user’s contacts. Once again, age homophily in the mobile call graph is biased by the age imbalance of the dataset. While the older generations do tend to have more older contacts and fewer younger friends, as seen in (c) and (d), younger generations, depicted in (a) and (b), also have a majority of middle aged or senior labeled contacts. That being said, it is visible that the percentage of younger contacts is smaller for older users, in (c) and (d), and, more importantly, that the percentage of contacts of a given age is bigger for users of that same age in detriment of others. As an example, the percentage of young friends is bigger in (a) than it is in (b), (c) or (d) and the same happens for the other age labels.

4.7 Relative frequency of communication among gender pairs. It stands out that females usually engage in more communications with other females than males do with other males. Notably, the frequency of communications between mobile phone users of different gender is also significant, possibly denoting social relationships of conjugal character.

4.8 Relative frequency of communication among age pairs. It comes as no surprise that users labeled as young engage in more frequent communications with their similar aged peers. While older generations communicate less frequently among themselves, there is an interesting tendency for young adults to communicate more frequently with seniors, when compared to the remaining cross-generation combinations. There is a high probability that users in those age gaps share relationships of parental character, which is likely to justify higher communication frequencies.

4.9 Percentage of users in different labeled triads in the mobile call graph by gender. (a) shows that of the females who are included in labeled social triads in the mobile call graph, most of them are connected in a triad with another female and a male. Contrarily, males are more repeatedly included in triads with all male members (b).

4.10 Demographic distributions among labeled triads in the mobile call graph. For both females (a) and males (b), older generations are the most frequent elements of triads in the mobile call graph, which suggests they have closer relationships with their contacts.
4.11 Network metrics of mobile phone users by gender and age. As opposed to what was found in [2], in this mobile call graph, younger generations have smaller degree (a), than older generations, which suggest the latter have a higher tendency to have more mobile phone contacts. The neighbor average degree assumes a completely different pattern for males and females, possibly affected by the higher percentage of males users in the network. As for clustering coefficient, (c) shows slightly higher values for younger generations and embeddedness (d) is relatively uniform for all demographic groups. Nothing concrete can be concluded about the tendency of having closer relationships for each gender and age group. The network metrics were calculated for a mobile call graph constructed from the mobile call dataset as defined in chapter 4 Model and Inference.

5.1 Illustration of a two-dimensional prediction space partition and the corresponding decision tree. There are five different possible prediction spaces each characterized by the conditions on variables $x_1$ and $x_2$ through parameters $\theta_1, \theta_2, \theta_3, \theta_4$.

5.2 Confusion matrix. An evaluation metric commonly used in statistical learning to evaluate the performance of prediction algorithms. The rows amount to instances in each predicted class, while the columns refer to the instances of the actual class. Of the actual positive observations the ones classified as positive are the True Positives (TP), while the ones classified as negatives are the False Negatives (FN). Of the actual negative instances, the ones classified as positive are False Positives (FP) and the ones classified as negatives are the True Negatives (TN).

5.3 Confusion matrices for gender prediction with the three models using 4 fold cross validation. Considering the imbalance of the test set, it would be desirable that the algorithms predicted correctly a higher number of males (TN) than females (TP), which happened for both Decision trees and RF, but not for GBT.

5.4 Confusion matrices for gender prediction with the three models using 10 fold cross validation. As in the case of 4 fold, both Decision trees and RF predicted correctly a higher number of males (TN) than females (TP), while GBT presented similar numbers.

5.5 Variable importance for gender prediction with Decision trees method using 4 fold and 10 fold cross validation. Network related features such as entropy of sms contacts, total number of triads and clustering coefficient appear among the five most important prediction variables.

5.6 Variable importance for gender prediction with GBT method using 4 fold and 10 fold cross validation. In the case of GBT, eigen centrality and closeness centrality appear among the more important variables, but are outnumbered by communication features.

5.7 Variable importance for gender prediction with RF method using 4 fold and 10 fold cross validation. For RF, eigen centrality and closeness centrality are among the most important predictive variables.
5.8 Confusion matrices for age prediction with the three models using 4 fold cross validation. Consistently the algorithms predicted correctly more instances of classes 3 and 4, which are the majority classes.

5.9 Confusion matrices for age prediction with the three models using 10 fold cross validation. As in the case of 4 fold, all algorithms predicted correctly more instances of classes 3 and 4 than 1 and 2.

5.10 Variable importance for age prediction with Decision trees method using 4 fold and 10 fold cross validation. Eigen centrality and closeness centrality are among the most important variables.

5.11 Variable importance for age prediction with GBT method using 4 fold and 10 fold cross validation. Like in the case of Decision trees, eigen centrality and closeness centrality are among the most important variables.

5.12 Variable importance for age prediction with RF method using 4 fold and 10 fold cross validation. Once more eigen centrality and closeness centrality play an important role in age prediction with tree-based classifiers.

5.13 Performance analysis for gender prediction with three tree-based methods, using 10 folds. A comparison between tests with network related features and without them is presented. There an accuracy improvement for all three methods, when predicting age considering network features. However, it is still a marginal improvement, which may further confirm the unsuitableness of such simple classification method for leveraging the network structure when inferring gender from call graphs.

5.14 Performance analysis for age prediction with three tree-based methods, using 10 folds. A comparison between tests with network related features and without them is presented. As a matter of fact, accuracy improvements are even less discernable in the case of age prediction using tree-based classifiers. There is a clear lack of capability to capture the network structure when inferring age from call graphs using classification methods.

A.1 Example of a small social network represented by a mobile call graph $G = (V, E)$. This network is composed by $|V| = 9$ users, connected among each other through $|E| = 14$ edges, conditioned upon the CDRs. Users $i$ and $j$ are connected in the graph if they communicated bidirectionally – either calls longer than 5 seconds or sms – at least once in 6 months.
A.2 Template model. The PGM is represented as an undirected graph, comprising two different kinds of random variables associated to each user $i$: $W_i$ represents the class and $X_i$ represents the set of features. For each $X_i$, there is an emission probability distribution $P(X_i)$. In terms of clique potentials, there exist unary potentials $P(W_i, X_i)$ and pairwise potentials $P(W_i, W_j)$. Since the random variable semantics are repeated for all users, a template probabilistic model is used for each user $i$ and the connections between $W_i$ and $W_j$ variable nodes are defined according to the connections between users $i$ and $j$ in the mobile call graph.

A.3 Representation of the model as a Markov Random Field. The independence network consists of an undirected graph $G = (V, E)$ whose nodes $Y \in V$ are the random variables $Y = W_1, \ldots, W_n \cup X_1, \ldots, X_n$. It defines the independencies between random variables and encodes the joint probability $P_\Phi(Y)$ as a product of factors $\Phi = \{\phi_1(D_1), \ldots, \phi_k(D_k)\}$, where $D_i \subseteq Y$.

A.4 Formal representation of the model as a Conditional Random Field. The independence network is represented by a partially directed graph $G = (V, E)$ whose nodes $Y \in V$ are given by $Y = W_1, \ldots, W_n \cup X_1, \ldots, X_n$. The distribution encoded considers a set of factors $\Phi = \{\phi_1(D_1), \ldots, \phi_k(D_k)\}$, where $D_i \not\subseteq X_i$. All $X_i$'s are assumed to always be observed, hence being depicted in grey.

A.5 Simplified representation of the model as a Conditional Random Field. The independence network is represented by an undirected graph $G = (V, E)$ and encodes the probability distribution of the target variables $W_i$, conditioned upon the observed variables $X_i$.

A.6 Spanning tree $T$ of graph $G$. $T$ is a spanning tree of $G$ if it includes all nodes in $G$ and the minimum number of edges to ensure graph connectivity without loops. The dashed edges belong to $G$ but do not belong to $T$.

A.7 Illustration of the method proposed in [4]. When breaking an edge in the graph $G$ to construct a spanning tree $T$, one should still take into account the contribution that the connection had for each of the end point nodes. That being said, node A will receive a contribution $\alpha$ and node B a contribution $\beta$, that respectively translates the connection between node A and node B, from the point of view of each node.

A.8 Scheme of the proposed algorithm. Weights are initialized for each edge of graph $G$ and from there an MST in constructed. Iteratively, inference is performed on the MST, after which the predictions are used to update the edge weights and create a new MST. A stopping condition is defined for when the predictions have not changed from the previous iteration.
List of Tables

3.1 Summary of network sampling methods: minimum assumption, sampling objective and properties preserved. FA - full access to network; PA - partial access to network; GC - graph connectivity; PE - property estimation; PP - property preservation; DD - degree distribution; CC - clustering coefficient, CM - centrality measures .......................... 27

5.1 Sensitivity, specificity and accuracy values for gender prediction with the three models, using 4 fold and 10 fold cross validation .................................................. 49

5.2 Sensitivity and accuracy values for age prediction with the three models, using 4 fold cross validation ................................................................. 51

5.3 Sensitivity and accuracy values for age prediction with the three models, using 10 fold cross validation ............................................................... 52
Contents

1.1 Motivation .......................................................... 2
1.2 State of The Art .................................................... 3
1.3 Thesis Outline ..................................................... 14
1.1 Motivation

Following the big digital revolution of the late twentieth century, we entered the Information Age. This is a period in which production, once industrial, became based on information and computerization. Companies understood that they must extend their focus from the product and encompass their customers. For that purpose, they realized there is a new form of capital for businesses: data.

The concept of Big Data was introduced as large amounts of data became available with the arise of social networks, the development of mobile smart devices and sensors, and the increased interest on the Internet of Things. Companies can leverage from the massive quantities of information to regain a competitive advantage by knowing their clients’ complete profile and behavioral patterns. However, it is common to have samples that are not representative enough. Take the example of online social networks: even though they are becoming increasingly popular, they are still mostly used by younger generations in developed countries.

One technology that has become ubiquitous is the mobile phone. For 2017, the International Telecommunications Union estimated 103,5 mobile-cellular telephone subscriptions per 100 inhabitants [5]. This means that telecommunications operators have in their hands large amounts of information that, more or less, span the entire population. More so, mobile phones record significant behavioral information, as people use them for their personal and professional communications. This information comes essentially in the format of Call Detail Records (CDRs). Logical charging functions deployed within the core network of the telecommunications system collect network resource usage per customer, allowing the construction of CDRs. The records generally contain identification of the end users, direction of the communication, start timestamp, duration of the communication and information of the base station that carried each voice, text and data traffic. These datasets have been of great interest to many different academic studies, entailing social, mobility and network analyses.

From CDRs one may construct call graphs, in which the nodes represent users and the edges represent communications between them. In the literature, one can add an edge under different conditions and the graphs can be directed or undirected, accordingly. Network science provides a variety of metrics that allow the study of graphs. Among other features, it might be relevant to analyze the node degree – the number of connection to a given node – and its distribution over the network.

Analysis of call graphs can give great insights about the patterns of users’ communications and how they are related to their demographics. Particularly, demographic properties such as age and gender influence the mobile phone users’ preferences and behaviors. The analysis of mobile traffic data “allows an operator to understand the behavior of customers, their calling patterns and habits, and thus to formulate adequate targeted offers.” [6] Moreover, models for predicting personality using phone-based metrics are being developed [7], while studies such as [8] acknowledge that “knowing the user’s personality can be a strategic advantage for the design of adaptive and personalized user interfaces.”

Nevertheless, demographic information in not always disclosed by clients, even though it is extremely relevant for the companies. A prepaid subscription, for instance, can be easily purchased
without the need of filling in any contract. Therefore, and since telecommunication operators do not have access to each user’s age and gender, they can try to predict it by capitalizing on the data available to them in the CDRs.

1.2 State of The Art

Inference problems are often solved using capabilities provided by fields as data science and machine learning, whilst relying on concepts of probability and graph theory.

Machine Learning (ML) gives computers the ability to learn from data. For example, algorithms learn the relationship between inputs and outputs in a dataset - where, given certain features, the expected outcomes to the problem are known – a labeled dataset. Fitting this knowledge, and under the assumption that the underlying probability distribution on the data does not change, it is possible to predict the outputs for an unlabeled dataset. In the case of supervised learning, labels are used to learn mathematical models for the data. Otherwise, an unsupervised approach is to be adopted. ML problems can be formulated as regression problems – where results are predicted within a continuous set of outputs – or classification problems – where the results predicted are discretized in classes. There can also be the case in which the target variable is not known – one would be in the context of unsupervised learning – and so the objective is to derive any structure from a dataset. Examples of this are clustering problems, where the data is clustered in different classes or categories, based on the features derived from the input data. Another example would be association problems, where the objective is to define rules that explain the relations among different variables within the dataset.

A typical ML problem revolves around a dataset that is divided into three different subsets: the training set, the cross validation set and the test set. The first one is used to fit a model to the data, while the cross validation set is used to select the better performing algorithm and minimize its error outside of the training dataset. Finally, the test set is used to estimate the accuracy of the selected algorithm’s predictions. When dealing with an inference problem the idea is to first try to model it with a simple hypothesis. There are some frequently used ML modeling frameworks, such as Linear Regression, Logistic Regression, Neural Networks or Support Vector Machines, and they all entail the minimization of a cost function. The parameters of the cost function are a critical aspect, since they will determine the accuracy of the predictions. They are learned, using specific optimization methods, such as Gradient Descent, to fit the training set. Once the cost function is parametrized, the cross validation set should be used to test the model. There are issues associated to the fitting of the parameters, which can affect the accuracy of the prediction model. A biased model is one where the data has been underfitted in such ways that some of the relevant relations between features are ignored. Contrarily, a model with high variance is very sensitive to minor fluctuations in the training set or, in other words, it is overfitting the training set, and, therefore, is likely to perform poorly in another dataset. Bias and variance can be diagnosed through plotting the error found in each dataset as a function of its size and the tradeoff model is depicted in Figure 1.1.

A high bias model would benefit from adding new features, but it might not benefit from getting
Figure 1.1: Bias vs Variance tradeoff. Lower complexity models are prone to higher prediction errors, suffering from high bias. This means that the model failed to capture major features of the training data and so, does not allow for accurate predictions on other datasets. On the other hand, a higher complexity model tends to have high variance, which signifies an overfitting to the training set (therefore obtaining minor prediction errors), and consequent poor performance in a different data sample. A good model will have to assume a compromise between complexity and prediction error, as depicted in the graphic. source: [1]

more data. As for high variance models, getting more training data is likely to help, as well as reducing the set of features. Often enough, the training examples have too big of a dimension and so dimensionality reduction can come in handy. The Principal Component Analysis, or PCA, is a frequently used method that tries to find a lower dimensional space onto which to project the data. Plotting learning curves can, then, help understand how to improve the initial algorithm. Additionally, Precision and Recall are two useful metrics to evaluate the performance of a ML model.

In the context of predicting mobile users’ demographics through analysis of call graphs, ML provides relevant tools. Inferring mobile phone users’ age and gender is a classification problem: given a set of features for each user (information retrieved from the CDRs and adequately preprocessed), it is possible to classify that user into male or female and belonging to a specific age class (e.g. young, young adult, middle aged or senior).

Another relevant framework for this problem are Probabilistic Graphical Models (PGM), inasmuch as this graph based representation provides a compact way of encoding a complex probabilistic distribution over a high-dimensional domain. In other words, it allows the representation of joint multivariate distributions over a large number of random variables that interact with each other. In a PGM, the graph is composed of nodes that represent the random variables and edges reproducing the direct probabilistic interactions among them. Its advantage lies in the fact that “it often allows the distributions to be written down tractably, even in cases where the explicit representation of the joint distribution is astronomically large”[9], making it accessible and more transparent for human understanding.

A PGM can be represented in terms of one of two classes of graphical representations of probabilistic distributions: Bayesian Networks or Markov Networks. The former use directed graphs, while the latter use undirected graphs. As seen in Figure [1,2] in a BN the graph representation is denoting a directed independence among variables, therefore encoding conditional probabilities. On the other hand, in a MN dependencies are not directed and one is encoding the joint probabilities of the random variables. When interpreting these independence networks, one can assume one of two
Figure 1.2: Example of a Bayesian Network (a) and a Markov Network (b). The graphical representation can be used to assess two different perspectives of the probabilistic graphical model. On the one hand, one can see the independencies induced from the graph: from (a), $(B \perp C|A)$, $(E \perp A|B,C)$, $(D \perp C,E|B)$, $(D \perp E|B)$, and from (b), $(A \perp D|B,C)$, $(B \perp C|A,D)$, where $\perp$ stands for conditional independence. On the other hand, the graph translates the factorization of the distribution such that $P(A,B,C,D,E) = P(A)P(B|A)P(C|A)P(E|B,C)P(D|B)$ in the case of the BN in (a) and $P(A,B,C,D) = \frac{1}{Z} \phi_1(A,B)\phi_2(B,C)\phi_3(C,D)\phi_4(D,A)$ in the case of the MN in (b).

perspectives. Firstly, one can see the graph as a compact way of representing the independencies among the random variables that hold in the probabilistic distribution. Secondly, the graph can define the skeleton for a compact representation of a high-dimensional distribution, in the sense that instead of encoding the probability of every possible assignment to all random variables of the domain, it is possible to factorize the distribution into smaller factors, each one over a smaller set of variables. Both representations provide the duality of independencies and factorization, but they differ in the set of independencies they can encode and in the factorization of the distribution that they induce.\[9\] Some probabilistic models may have a repetition pattern either in time or in terms of the random variable definition and can also be represented by graphical models, leveraging from such repetition. Temporal models can be used when trying to model dynamic settings, where the focus is to understand the evolution of the state of the world as time progresses. An even broader framework is that of Template-based models, in which a template is instantiated several times within the model, useful when encoding models with random variables that share the same space and semantics.

Another advantage of PGMs is that this sort of graphical structure allows the probabilistic distribution to be used effectively for inference. When modelling a specific reality, one often aims to make queries about the distribution of the model. In fact, there exist various algorithms for computing the probabilities of some random variables, given evidence about the others. As shown in [9], “the structure of the network, both the conditional independence assertions it makes and the associated factorization of the joint distribution, is critical to the ability to perform inference effectively, allowing tractable inference even in complex networks”. Depending on the inference task, exact or approximate algorithms can be used. When aiming to make conditional probability queries, in which the purpose is to compute the probability of a subset of variables $Y$ given some evidence $E = e$ about other variables in the model, $P(Y|E = e)$, algorithms such as Variable Elimination provide exact inference. Others such as Message Passing algorithm and Random Sampling Instantiation algorithms may only provide approximate inference in certain cases. Besides, one may want to use a Maximum a Posteriori (MAP) approach to inference: given evidence for a set of variables $E = e$, the objective is to determine which is the assignment $Y = y$ to remaining variables in the model (for which evidence is not provided) that
maximizes the conditional probability $P(Y|E = e)$ \[^9\]. Indeed, there may not be a unique solution to a MAP assignment and so there may be more than one possible solution to this inference problem. To solve a MAP assignment problem, Variable Elimination and Message Passing algorithms can also be used. However, because MAP assignments are intrinsically an optimization problem, it is also possible to rely on techniques that come from optimization such as integer programing, graph-cut or combinatorial search methods.

Additionally, there might be the case in which one has access to sets of data with examples from past experiences and wishes to learn a probabilistic model to describe such reality. PGMs facilitate the process of learning from data with the objective of later completing varied inference tasks. There are different learning methods and their merits can only be evaluated taking into consideration the specific inference goal. In practice, the data that is available is often not enough to construct a ground-truth model and so one must aim to find the best possible approximation. Naturally, the notion of best will heavily depend on the goal and different models will encompass different trade-offs.

A class of probabilistic graphical models that is extremely popular in various areas is discrete Markov Random Fields (MRF), for which optimization methods have been the focus of many recent research studies. MRFs have had great success in fields ranging from computer vision or digital communications to statistical physics and machine learning. Optimization using such models often corresponds to the MAP estimation problem and is usually “directly described as an energy minimization task” \[^10\]. As such, various algorithms for discrete energy minimization on MRFs can be found in the literature.

In \[^11\], Wainwright \textit{et al.} present a technique for optimization in MRFs – tree-reweighted max-product message passing (TRW). The TRW method is a message-passing based method, which aims to find a maximization of the lower bound of the energy and that presents itself as a rival to widely deployed methods like Belief Propagation message passing or graph-cut based methods. The energy minimization problem is usually defined in the form of an integer program, which is then approximated using a linear programming relaxation. Based on the assumption that the standard LP relaxation provides a good approximation, TRW methods hope to obtain an approximately optimal solution for the MAP estimation problem. Instead of minimizing the linear programming approximation directly, these methods focus on solving the dual function of that relaxation, that way aiming to find a maximum of the lower bound of the optimal MRF energy. Yet, TRW algorithms as proposed in \[^11\] are not guaranteed to increase the lower bound on the energy and may even decrease it, as observed by Kolmogorov in \[^12\]. In addition, the algorithms do not converge.

Taking advantage of powerful ML inference methods and PGM frameworks, telecommunication operators can get to know their customers a little better without overstepping privacy issues and, thus, provide a personalized and effective service. As a matter of fact, several studies have been conducted on the problem of inferring users’ demographics from mobile call graphs.

To further explain the state of the art methods it is important to first clarify a few concepts. The first concept is that of an ego network, illustrated in Figure \[^1.3\] In a mobile call graph, it consists of a user, as a central node, and his first-hop neighbors, i.e. nodes directly connected to him.
Figure 1.3: Example of an ego network. A network is composed of vertices, the circles, and edges, the lines connecting them. The central white vertex represents the ego of the network, i.e., the central node. Depicted as dark circles are the immediately connected vertices, the ego’s neighbors. An ego network further includes the connections among them.

Figure 1.4: Example of a factor graph. The \( X_i \) nodes are variable nodes, while the \( f_i \) are factor nodes. Connection between pairs of variable nodes are made through factor nodes, that represent the mathematical dependence of the variables. Factors can be related to a single variable, as is the case of \( f_1 \), and it is also possible to have more than one factor respective to the same pair of variables, as happens with \( X_1 \) and \( X_2 \) through \( f_2 \) and \( f_3 \).

An additional relevant formulation is the one of factor graph, which is the underlying representation for inference in graphical models. As mentioned before, PGMs use graphical representations to express the conditional independencies between random variables. A factor graph illustrates the factorization of a function, permitting a global function of several variables to be expressed as a product of factors over subsets of those variables. It is bipartite, since it includes two different types of nodes: variable nodes – which can be evidence variables, when their value is known, or query variables, when their value should be predicted – and factor nodes – defining the relations between the variables –, as seen in Figure 1.4. Factor graphs are particularly useful in the context of probabilistic inference, which refers to deriving the probability of one or more random variables taking a specific value or set of values. Given that they can represent the factorization of a probability distribution function, factor graphs allow more efficient computations of, for example, marginal probabilities, when combined with message-passing algorithms. A commonly used message-passing algorithm on factor graphs is the sum-product algorithm, where messages are conceptually computed in the nodes and passed along the edges. Inference can be efficiently and effectively performed on tree-structured graphs with message passing algorithms. In practical applications, however, most graphs have loops and exact inference is not feasible. One available approach is to use algorithms such as Loopy Belief Propagation, which allow approximate inference.[13]

One of the state of the art approaches, presented by Wang et al. [14], explored the homophily patterns within mobile call graphs. Homophily is defined as the tendency that humans have to relate to others with similar demographic and behavioral characteristics. The study was performed for three areas of demographic information: home location, income level and age group. For this thesis, only
the latter will be taken into account. In this specific context, they begin by confirming that there exists age homophily in the call network. They proceed by identifying which features imply higher levels of homophily between users and, finally, they try to understand how homophily in the call graph could be exploited to infer a user’s demographics using the demographic properties of his/her friends.

The dataset is composed by CDRs with call information only, collected over a 1 month period in the USA in 2010. Of the 93 million callers that compose the dataset, approximately, 20 million have associated demographic information. When constructing a call graph, all users are represented as nodes, and edges between two users translate at least one bidirectional communication. They use three different undirected, unweighted graphs associated to different time slices – weekdays, nights and weekends, and a joint graph. They do so because previous work has confirmed varying behaviors along these different time frames. Users will be classified according to four age groups: 18-30, 31-45, 46-65, and over 66, accounting for 14.94%, 40.31%, 33.62% and 11.12% of the labeled population, respectively. Features are attributed to each edge of each graph and entail number of calls, average call duration, degree difference and number of common connections.

After verifying that all subscribers have more than 50% chance of calling others of the same age group, thus confirming age homophily, they proceed by identifying the features that are more representative of such homophily. For that purpose, they use a linear regression model, concluding that the least significant features are the number of night-weekend calls – indicating that people make more calls to those of the same age group during weekdays – and the number of night-weekend common friends – showing that subscribers with more weekday common friends have higher probability of being of the same age.

A random set of 1 million subscribers is used for experiments, of which 80% is the training set and 20% is the test set. For age group inference, two algorithms are tested. The first corresponds to a Majority Vote (MAJ), which classifies a subscriber with the most dominant age group of its neighborhood. RANK, the second algorithm, uses the most relevant features of each edge to assess the level of similarity between the connecting users and classifies a target user according to its most similar neighbor. Both algorithms perform better in the weekday graph, achieving accuracies of 78% and 77%, respectively, further emphasizing that the weekday behavior is the most representative for age prediction. Using the complete graph, they conclude that RANK outperforms MAJ on low-degree subscribers and the opposite happens for high-degree subscribers.

One fundamental parameter in this study should be the stopping criteria for the MAJ algorithm, although it is not mentioned. If it is to be run through the whole graph, a final state would mean all nodes are classified with the same age group. As for the RANK algorithm, the previous concern does not apply. The most similar neighbor is selected based on their features, thus converging to a final state that is unaffected by the selections made for the other nodes in the network. Even though homophily in call graphs, when proven, can, in fact, be explored for demographic prediction, this first approach seems to be a rather limiting one. Not only it might be possible that one communicates with people of predominantly different ages, but also looking only to the direct neighborhood of a node does not allow influence to spread over the network, in the sense that second and higher-hop
neighbors are disregarded as possible influences of one's age.

In order to incorporate spreading of influence throughout the call network, Brea et al. [15] propose a different method for age prediction, in which the information arising from second or higher-hop neighbors is also taken into account. After confirming, once more, the strong age homophily in the call network, the objective is now to find the probabilities of a node belonging to each age group, based on the topological relations in the mobile call graph. The dataset includes CDRs with information about calls and SMS (short message service) for around 93 million mobile phone users over a 3 months period from a mobile operator in Mexico. To construct a call graph, edges are added between two nodes that communicated at least once, either by call or by SMS. In the joint graph, including both operators' clients and other users, only a subset of the client nodes is labeled with information about age - referred to as seeds. Of that, 75% is used to generate the predictions and 25% is used to evaluate the performance of the algorithm. They explore two different undirected, unweighted graphs - the joint graph, consisting of approximately 70 million users, after data pruning; and one composed by clients only, with around 3 million nodes. The age partitions correspond to under 25, 25-34, 35-49 and above 50. As for the topological metrics used for inference, they used number of seeds in a node's neighborhood, topological distance of a node to its closest seed and node degree. The proposed method entails a Reaction-Diffusion algorithm: each node is attributed a vector with the probabilities of belonging to each age group, independently, and the vector is updated iteratively using an algorithm that considers two terms. The reaction term represents the initial state of the probabilities – complete certainty for seed nodes and equal probability of belonging to each age group for the remaining. The diffusion term consists of a mean field of the node's neighbors probabilities. To reduce the diffusion noise, they consider the mean field generated only by neighbors that have received information from seeds. After some iterations each node is assigned the age group with the highest probability. Collapsing all probability vectors into one age group in the end has the drawback of not taking into account the remaining results for the other nodes. It is recommended by the authors that some external constraints should be included, such as preserving the seed's age distribution in the entire network. For that purpose, they propose a method, which they call Population Pyramid Scaling or PPS. Running the algorithm for 30 iterations on a network where the maximum distance from a node to a seed is 20, convergence is guaranteed, and the best performance was achieved for the 35-49 age group, 52.3%. The overall performance on the validation set was of 46.6% for the joint graph without constraints, and 46.5% using the PPS algorithm – showing that the method without PPS was able to preserve the demographic distribution of the seed set. For the clients only graph, the achieved performance was of 49.9%. All cases performed better than a random guessing approach (25%) or assigning the most probable age group to all nodes (36%). The features are then used to understand in which conditions does the algorithm perform best. They verify that the highest performance of hits is 66% for nodes with four seeds in their neighborhood and that the algorithm performs best for high degree nodes (achieving peak performance for nodes with degree equal to 10). They observe that almost all nodes are at a distance smaller than 4 of the closest seed, meaning that with just 4 iterations the seeds' information would have spread throughout the whole network.
Optimizing over all three metrics and choosing an adequate subset of 20,050 nodes, it was possible to achieve a performance of 62%.

Sarraute et al. [16] have conducted a consequent study on the exact same dataset that combines the graph based inference approach of the previous Reaction-Diffusion method with classical Machine Learning techniques based on node features. They further extended the analysis by including gender prediction. For the clients of the mobile operator, they gathered behavioral features: number of calls, duration of calls, number of SMS, number of contact days, in-degree, out-degree and degree. For the first 3 features, 12 variables were defined, distinguishing between incoming, outgoing and all and computing for different time slices - weekday light, week night, weekend and total. As for number of contact days, this feature was unfolded into 6 variables, distinguishing between calls and SMS, and between incoming, outgoing and all. The logarithmic version of these variables was considered, so as to mitigate the impact of the heavy tail distributions of some of them on the Machine Learning algorithms. They performed PCA in order to understand which were the most relevant features, concluding that the most important ones included the logarithmic version of the total number of calls, total duration of calls and total number of SMS, after which outgoing variables were preferred. Once again, the objective is to collapse the probability vectors, using the PPS algorithm to preserve the seeds’ age group distribution. When using the previously mentioned constraining algorithm, predictions are made for a fraction $q$ of the nodes which where best classified with the unconstrained methods.

For gender prediction they used Naive Bayes, Logistic Regression, Linear SVM, Linear Discriminant Analysis and Quadratic Discriminant Analysis. The best results were obtained with Linear SVM and Logistic Regression. Using PPS, it is clear that there is a trade-off between precision and coverage (determined by $q$, the fraction of nodes to which predictions are made for): the highest accuracy of 81.4% was obtained for 12.5% of the users ($q = 1/8$), while predicting for the whole set ($q = 1$) allowed for an accuracy of 66.3%. As for age prediction, they tackled the problem in two different ways. Firstly, they made a prediction based on node attributes, using Multinomial Logistic, which is a generalization of Logistic Regression for multiple class classification. The model was overfitting the categories with higher frequencies and achieved a maximum of 52.7% accuracy for the best dataset configuration ($q = 1/8$) and 36.9% for the whole set ($q = 1$). Alternatively, they used a prediction method based on the network topology, taking advantage of the previously developed Reaction-Diffusion method. On its own, this algorithm performed better than the Multinomial Logistic achieving 62.3% in the most favorable configuration and 43.4% in the least favourable one. It also outperformed a combination of the two methods, using the Multinomial Logistic results as initial states for the Reaction-Diffusion algorithm.

From these two studies it was possible to conclude that a diffusion term has a relevant part, when predicting age from call graphs. It seems informative to look into the network as a whole as opposed to looking only at each node’s ego network, when predicting demographics. Iterating over a probability vector instead of a single value, allows the system to evolve as a probability state over the whole network. This suggests that a Probabilistic Graphical representation might also suit this inference problem.
A rather interesting approach to the problem is proposed by Dong et al. [2], where age and gender are predicted simultaneously. The novelty being that, not only the relations between age/gender and the communication features are taken into account, but also the interrelation between the two demographics is explored.

Studying the underlying social strategies of call graphs, they found evidence of strong homophily for both age and gender. Additionally, they observed interesting cross-generation communication patterns, that, for example, relate to parent and children relationships, as well as different demographic dynamics spanning one’s lifetime, according to gender. On top of these discoveries, a double dependent-variable prediction method is defined, meaning that they aim to determine the joint probability of a user being of a certain gender and belonging to a certain age group, given its communication patterns within the mobile call graph. To do so, they adopt a factor graph representation, defining three types of factors - attribute factor, that represents the correlation between the user’s demographics and the communication features retrieved from the graph; dyadic factor, that captures the relation between user’s demographics and the ones of his direct relationships in the ego network; and a triadic factor, modeling the correlation of users’ demographics within triadic relationships in the ego network. An illustration of the propose model is presented in Figure 1.5. Model features include typical network metrics, such as degree centrality, neighbor connectivity, triadic closure, embeddedness and other friend and circle specific features as, for instance, number of different gender triads in which the user is included.

The data set was collected in 2 months of 2018 and includes around 7 million users from an anonymous country. Call and SMS data is treated separately and so two undirected and weighted graphs are constructed, where an edge represent at least one bidirectional communication between the two nodes. Since the age group between 20 and 55 years old is overrepresented, they partitioned the dataset into four age groups of 18-24, 25-34, 35-49, and over 50. Exploiting the factor graph representation, the joint conditional probability of the variables age and gender, given the features

![Figure 1.5: Illustration of the graphical model proposed by Dong et al., where $y_i$ is gender, $z_i$ is age and $x_i$ is the feature vector of user $i$. Three different kinds of factors were defined: $f(y_i, z_i, x_i)$ represents the relation between demographic variables and the set of features of user $i$, $g(y_i, z_i, y_j, z_j)$ represent the dyadic relation between users $i$ and $j$ and $h(y_i, z_i, y_j, z_j, y_k, z_k)$ refers to the triadic relation among users $i$, $j$ and $k$. Source: [2]![](https://example.com/figure1.5.png)](https://example.com/figure1.5.png)
redeemed from the network structure, is written in terms of the product of the three proposed factors. The factors are parameterized and such model parameters are learned by maximizing in the training set (50% of the dataset used) the log-likelihood of the objective function – a combination of all factors representing the joint conditional probability of a user being of a certain age and gender, given its features –, using the Gradient Descent technique. The aforementioned process entails the computation of the marginal probability estimates for the unknown variables age and gender, by means of Loopy Belief Propagation. Once the model is parametrized, prediction is made by maximizing the objective function itself.

The suggested method achieved 80% accuracy for predicting gender and 73% for predicting age. It outperformed other methods used for comparison, namely Logistic Regression, Support Vector Machines, Naive Bayes, Random Forest, Bagging, Gaussian Radial Basis Function Neural Networks and a simplified Factor Graph model (for independent prediction of age and gender). Distributed learning is also considered for scalability, adopting a Message Passing Interface framework to split the network into smaller parts and learn the model parameters on separate processors.

Factor instantiation is particularly decisive for the performance of the algorithm and the rationale behind the adopted expressions is left unclear. The same can be said for factor initialization. Furthermore, the feature definition seems to be restricted to one's ego network, and could possibly be enriched by considering second or higher-hop neighbors.

Analyzing all four state of the art approaches, it is clear that the latter is the most sophisticated one and yields the best results in terms of predictive accuracy, possibly being a good starting point for tackling this concrete inference problem. Nevertheless, it leaves space for improvements.

There is additional literature related to the study of social behaviors and demographics properties based on mobile call information, which is worth mentioning.

One interesting study was conducted by Frias et al [17], with the aim of characterizing and automatically identifying the gender of mobile phone users using the behavioral, social and mobility information presented in CDRs. It is pointed as being of great interest in the particular case of developing economies for segmented cell phone-based services, but with the potential of being extended to other contexts. The study entails the understanding of the role of gender in mobile phone usage and evaluating common ML methods for gender automatic identification.

CDRs collected over a 3 months period are used to collect information on around 10 thousand unique mobile phone numbers of an unidentified telecommunications operator. From that information, three sets of variables where defined for each user: behavioral, social and mobility variables. The first set includes number of incoming and outgoing calls, average duration of incoming and outgoing calls and expenses made by user over a period of \(d\) days. Furthermore, social variables include in and out-degree and network degree – the number of unique phone numbers that have either contacted or been contacted by each user. Finally, the mobility variable corresponds to the route distance, i.e. the distance travelled by each user in between consecutive calls. For each variable, the distribution over each gender type (female or male) was computed. Previous studies indicate that calling behavior typically assumes either pow-law, log-normal or exponential gender distributions, and the parameter
for all three distributions were fitted for each variable-gender pair. It became clear that outgoing and incoming variables have similar distributions. Using the Kolmogorov-Smirnov goodness-of-fit statistic, the distributions of number of calls and degree were best approximated by a power-law; the best fit to the distributions of average number of calls and route distance was the log-normal distribution; and as for the distribution of the expenses over each gender, it is best suited by an exponential distribution.

The general results indicate that the mobile phone usage levels are higher for females than for males, regarding both behavioral and social metrics. Performing t-test and a two-sample Kolmorov-Smirnov test for each pair of variables distributed over female and male populations, the conclusion was that the differences between female and male distributions are statistically significant.

With respect to gender prediction, the process begins with identifying a calling behavior for each user, based on the most statistically relevant features: number of calls and average duration of calls, both incoming and outgoing, as well as expenses and network degree. The set of unique users was, then, divided into training set, accounting for 70%, and test set, the remaining 30%. Using 60-fold cross validation, the SVM and Random Forest methods were tested, obtaining 54.2% and 56% accuracy, respectively. Since the improvement from random guessing (50% chance of predicting correctly), was marginal, a new method was proposed: using a semi-supervised clustering algorithm. Using the training set, the k-means algorithm allows distributing all user behaviors into k clusters, after which it is possible to classify each cluster according to a minimum threshold of female/male labels. Naturally, the higher the percentage requirement for classifying as female or male, the more precise will be the classification. Those clusters which do not comply with the requirements, do not get classified. A radius measure is computed, corresponding to the maximum distance of a user’s behavior point to the centroid of the classified cluster it belongs to. For each user behavior in the test set, the algorithm will find the closest labeled cluster and attribute its classification, provided that the distance of the behavior to the center of the cluster is smaller or equal to the radius of said cluster. Otherwise, the user gets classified with a do not know label. There is a clear trade-off between coverage and accuracy of the predictions. An accuracy of 80% was obtained with 3% coverage, using 50 clusters. But a 12% coverage implies a smaller 70% accuracy, for example.

The proposed method allows for the verification that, for all tested numbers of clusters, the labeled clusters that produced the highest accuracies were tagged as female, implying that there exists a group of females with above average mobile phone usage. That being said, gender prediction can be a powerful tool for segmented services that rely on mobile phone usage.

Another interesting analysis performed by Stoica et al. [18] confirms, yet again, the different communication behaviors associated with age and gender and goes as far as to say that the “gender of the receiver of phone calls is strongly correlated with the duration of the calls, while the gender of the caller is less important”; since they concluded that the callers tend to adapt their interactions to the gender of the destination of the communication.

On another note, the social brain hypothesis conjectures that humans have approximately 150 relationships at any given time, divided into layers of proximity. Dunbar suggests that the number of friends in a layer decreases as the emotional closeness increases, defining numbers of 5, 15, 50
and 150 for the cumulative amount of friends at each layer. Mac Carron et al. [19] studied to which extent can these layers of emotional closeness be identified within mobile communications. For that purpose, they assumed frequency of call as a good indicator of the strength of a relationship, which is consequently associated to emotional proximity.

Utilizing an European dataset collected over the year of 2007 with approximately 34.9 million users, the aim was to be able to cluster one’s friends into layers. Given the background of the social brain hypothesis and Dunbar’s numbers, a cut-off of 100 unique contacts was used and so, around 27 thousand users where analyzed. Three methods were applied for clustering. The first was the Jenks algorithm, which works similarly to the k-means algorithm. It divided almost half of the users into 4 to 5 clusters, with the number of friends belonging to each cluster approximated to the ones proposed by Dunbar. Secondly, the analysis was performed with a Gaussian Mixture Model and resulted in most users being split into 2 or 3 clusters. Yet again, the numbers of friends resembled the one’s in Dunbar’s outer layers. Lastly, a head/tail breaks algorithm was adopted, splitting the majority of users into 3 or 4 layers, with similar cumulative number of contacts.

Even though the number of clusters and consequent cumulative sizes of each layer vary both in the method used, and when compared with Dunbar’s prediction, relevant similarities were found. Applying different clustering algorithms to the dataset, “strong evidence for a layering structure” was discovered. Indeed, this study shows that the strength of human relations is observable through calling patterns present in mobile phone dataset, hence proving how important these datasets can be when studying social and demographic properties of users.

In conclusion, there already exists vast literature on the potential of analyzing mobile phone datasets for purposes ranging from academic to social and even business related. Some studies have, inclusively, tackled the issue of predicting demographic properties, such as age and gender, from CDRs. They achieve prediction accuracies of around 50% to 80% in the best configurations, using relatively low-complexity models. On such grounds, the relevance of this thesis is confirmed.

1.3 Thesis Outline

This thesis is divided into five main sections. In a first chapter the concept of mobile call graph is discussed. There is a brief definition of the broader concept of graph followed by an explanation of how it can be used to represent a social network composed by the mobile phone users in Section 2.1. In addition, in Section 2.2 some relevant properties from graph theory are outlined, since they permit interesting analyses of the mobile phone users’ behaviors as nodes of the graph. Prior to starting the analysis of the data, a brief remark is made on network sampling in Chapter 3. Since most real world networks are composed by millions of nodes, network sampling is a powerful tool that allows studying a smaller representation of large networks that would otherwise be computationally complex or even infeasible. The analysis of the data begins on Chapter 4. There is an introduction to exploratory data analysis in the context of a data science project in Section 4.1 followed by a presentation of the findings in Section 4.2. The latter entails the description of the mobile call dataset in Subsection
4.2.1 an assessment of the variables of the dataset in Subsection 4.2.2 and a presentation of the social strategies in Subsection 4.2.3. The final conclusions for the EDA process are summarized in Subsection 4.3. Once the EDA is exposed, a preliminary approach using classifiers is proposed in Chapter 5. A description of the inference problem can be found on Section 5.1 followed by Section 5.2 where the tested methods are exposed. The experiments setup and results analysis are included in Sections 5.3 and 5.4. A final section details some related work. In Chapter 6 conclusions are drawn and future work is proposed.
2 Mobile call graphs

Contents

2.1 Definition ................................................. 18
2.2 Properties .............................................. 19
Mobile call graphs are a fundamental structure to the representation of mobile phone users as a social network. This chapter begins with a definition of the broader concept of graph and details how it can be applied to the context of mobile phone users. In a second section, there is an overview of the properties stemming from graph theory and which will have an impact on the analysis of this network.

2.1 Definition

As it was previously mentioned, mobile phone communications are recorded in CDRs, detained by the telecommunications operators. These records hold the details of each communication – but never the content –, as well as identifiers of its intervenients, and can be paired with extra personal and contractual information that the operators might have on their subscribers. In order to understand the communication behaviors of users of mobile phones, one can try to observe them as a social network and convey the relations among them. In fact, call graphs are an extremely interesting way of representing such social network in a graphical way, particularly given the usefulness of various well-known metrics and analysis techniques stemming from graph theory.

A graph $G = (V, E)$ is a mathematical structure composed by vertices $V$ (also referred to as nodes) and connections between pairs of them, denoted by edges $E$. Edges in a graph can be undirected, representing a bidirectional relationship between the end nodes, or directed, representing a one-way relation from the starting node to the end node. They can also be weighted or unweighted, meaning that they can have a weight or cost associated to them or not, respectively. Figure 2.1 shows an example of an undirected and unweighted graph composed by 6 nodes and 6 edges. Even though many have deployed call graphs in the study of mobile communications, there is not a unified way of defining a call graph, but rather many different variations are found in the literature. Such variations arise from the different possibilities when defining an edge between two nodes.

When constructing call graphs, it is typical that the starting point is a mobile dataset, in which columns have information on a pair of users and the communication(s) between them. This kind
of dataset is composed by either the CDRs themselves or a processed and condensed version of them. It is, then, imperative that one defines what is commonly denoted as edge condition. It is, essentially, the rule that determines whether an edge between two nodes will appear in the graph or not, based on the observations on the mobile dataset. From graph theory, it is know that edges can assume many properties regarding orientation, weight and they can additionally be subject to filtering requirements[6], as will be explained further on.

The definition of the edge condition format was based on the literature. Both Wang et al. [14] and Dong ev al. [2] introduced an edge when there was at least one bidirectional call, the first within 1 month and the second within the 2 months observation period, resulting in undirected graphs. Conversely, Brea et al. [15] and Saurrate et al. [16] used an edge condition of at least one communication (call or sms), producing directed graphs. Stoica et al. [18], for instance, connects two vertices with an undirected edge if there has been at least one communication each month, adding to a total minimum of six calls within the observation period. With respected to the weight of edges, most state of the art studies disregard it, while there exist studies that do not. An example is the one conducted by Mota-hari et al. [20] on categorizing social ties from call data records, where to identify different strengths of relationships, it is useful to portray different edge weights.

That being said, all the previously mentioned literature was taken into account for this particular study. With respect to the graph creation, a function was created for the generation of a mobile call graph based on a dataset and a specific given edge condition. To define an edge condition five important parameters were identified: communication type, communication direction, limit on number of calls, limit on number of sms and limit on average duration of call. Consequently, the edge condition is represented by an array of five elements. The communication type is a binary variable that can assume the value 0 when there should be no distinction between a call and an sms, and 1 otherwise. Similarly, the communication direction is also a binary variable, being 0 whenever considering unidirectional communication, meaning that it flows one way only, and 1 when considering bidirectional communication, i.e., when a given user called another and was also called by that same person – or sent and received an sms. This is a significant parameter of an edge condition, since it determines whether the graph is directed (unidirectional communication) or undirected (bidirectional communication). As for the limit on number of calls and sms, both variables should be integers and represent the minimum number of communications (respectively, call or sms) that stands as a requirement for the introduction of an edge between two users. Finally, the limit on average duration of call is a self-explanatory float variable.

2.2 Properties

Call graphs are characterized using interesting metrics that allow for relevant analyses of the different characteristics of users and the way they communicate [8].

Firstly, there is graph connectivity, a basic concept of graph theory. A graph $G$ is said to be connected, when there are no unreachable nodes, that is, when all nodes that belong to $G$ are connected
to each other, direct or indirectly. Contrarily, when a graph $G$ is disconnected, there is at least one pair of nodes for which there is no path in $G$ that has those nodes as endpoints. For disconnected graphs, a connected component is defined as the maximal connected subgraph, and each vertex and edge belong to exactly one connected component. In the context of mobile call graphs, a fully connected graph would represent a fully connected social network, in which all members are connected directly or through mutual contacts. It is unrealistic to expect that the mobile call records of telecommunications operators would produce such a graph. Thus, it is common to analyze exclusively the largest connected component of a graph created from a mobile dataset.

Also related to paths between nodes in a graph, is the diameter metric. The diameter of a graph $G$ is defined as the maximum distance between any pair of nodes in $G$. In itself, the distance between two nodes is the number of edges in the shortest path connecting them, also known as the geodesic distance, as illustrated in Figure 2.1. There may be more than one shortest path between two nodes. In the case that there exists no path connecting two nodes, it can be concluded, that they belong to different connected components of the graph. Moreover, the diameter of a disconnected graph corresponds to the diameter of the largest connected component.

Thirdly, the degree distribution is a powerful concept. It corresponds to “the statistical distribution of the number of vertices connected by edges to a single other vertex. It conveys information about the basic structure of communications among mobile users.” [6] In directed call graphs, it is possible to distinguish between incoming edges – in-degree – and outgoing edges – out-degree –, thus preserving the relations between caller and callee. In undirected graphs, however, a single measure of degree includes all edges connecting to a node.

Both in the natural and virtual world, power-law distributions have become ubiquitous in describing a varied number of phenomena [21]. Indeed, many authors have fitted degree distributions of call graphs with power-laws. For the case of undirected graphs, Onnela et al. [22] fitted a power-law to the node degree distribution with an exponent of 8.4, whilst Lambiotte et al. [23] fitted the node degree distribution with a power-law with exponent of 5.0. Considering directed call graphs, Nanavati et al. [24] observed that both in-degree and out-degree distributions can be fitted by a power-law, with exponents taking slightly different values: ranging from 2.7 to 2.9, in the case of in-degree and from 1.5 to 2, for out-degree. Additionally, they concluded that there is strong correlation between both metrics, implying that users that tend to communicate with more people are also likely to be contacted by a large group of mobile phone users. Notwithstanding, the same relation is lost for the case of high in-degree nodes (eg. customer service numbers) and high out-degree nodes (eg. salesman). Doran et al. [25] have drawn similar conclusions, obtaining exponents of 3.41 and 2.63 for the fitting of the in-degree and out-degree, respectively. Interestingly, the node degree distribution in undirected graphs presented higher values for the power-law exponents, than those obtained for in-degree and out-degree, in directed graphs. Higher values for the exponents reveal a higher number of high-degree nodes. Therefore, the idea behind the previous comparison is to observe that there might exist less high-degree nodes in directed in directed graphs – where only mutual communications are considered – when compared to directed graphs – where one-way interactions are included. As
so, it appears that people tend to have fewer contacts for which communications are both made and received from the other user. In the case of this thesis, an exponent of 4.23 was obtained for the power law distribution fitted to the degree distribution of the mobile call graph (the construction of which will be detailed further on). The degree distribution, plotted in Figure 2.2, suggests that, as the power law exponent is relatively smaller than the ones found in the literature for undirected graphs, in this mobile network users do not tend to have many contacts.

Even though all previously mentioned authors were successful in fitting power-laws to degree distributions, they found that, in general, power-laws fitted well the tail of the degree distributions, but the same was not true for the head. Alternatively, [6] and particularly [21], propose the Double Pareto Log Normal (DPLN) distribution as yielding a good fit for the entire node degree range. A DPLN distribution arises from a combination of lognormal distributions as is thoroughly discussed in [26].

In addition, clustering coefficient is a measure of the degree to which graph nodes tend to cluster together. Sometimes referred to as transitivity, this concept is not a new idea in the study of interpersonal relations and is, for example, fundamental to Heider’s balance theory. In [27] the authors mention that “Although Heider’s theory was originally intended to refer only to cognitive structures of an individual person, […] the definition of balance may be used generally in describing configurations of many different sorts, such as communication networks […]”, and so the concept of transitivity applied to this particular network can be very interesting. In fact evidence suggests that in most real-world social networks, nodes tend to create tightly closed groups [28]. There exist two versions of this metric: global and local clustering coefficients. The first is based on triplets of nodes, which are structures of three nodes connected by either two – open triplet – or three – closed triplet – undirected ties. A triangle in a graph includes three closed triplets, centered in each node. Global clustering coefficient is then defined by the number of closed triplets, or three times the number of triangles, divided by the number of all triplets (open and closed) in a graph. As for the local clustering coefficient of a node, it defines how close are the node’s neighbors to being a clique, i.e., a complete graph. It is measured by the proportion of links between the vertices within the node’s neighborhood divided by the number of links that could possibly exist between them. When inferring user demographics from call graphs,
local clustering coefficient can be a relevant metric to assess the local connections among nodes and their immediate neighbors.

To finalize, there is one more pertinent property, emerging from network science: the centrality measure. There are, in truth, many variations of this measure that is aimed at identifying the importance of nodes within a graph. The most common is the already mentioned notion of degree, which accounts from the number of links incident upon a node. Furthermore, there is closeness centrality, measuring the average length of the shortest path between one node and all the others. The more central a node is, the closer it is to the remaining nodes. Moreover, betweenness centrality is a similar concept that traduces the number of times a node is located along the shortest path between any two other nodes. And finally, the eigenvector centrality is used to assign relative scores to nodes in a graph, based on the concept that connections to high-score nodes contribute more to one’s score than equal connections to low-scoring nodes. It is computed utilizing the adjacency matrix of a graph, which is a square matrix indicating which vertices are adjacent to each other in a graph. All in all, centrality measures are interesting inasmuch as one wants to identify which nodes are the most important and, thus, influential, in a network.

All in all, leveraging from graphical structure representations and graph theory concepts can allow for a better understanding of the mobile phone social network. Moreover, graph properties can be used as features in the estimation of demographic properties from call graphs. The degree is an example of a network metric that is used as a feature in all state of the art approaches[14], [15], [16] and [2]. Furthermore, centrality measures can also be relevant when trying to discover influential nodes in a network [29].
Network sampling
Most real world datasets often represent millions of users, resulting in extremely large networks. In order to analyze and conduct algorithm experiments for such networks it is useful to find a smaller representation through network sampling. This chapter describes the purposes and scenarios to which network sampling applies and presents a comparison of the different possible methods developed to date. Thenceforth, an explanation is given on the application of network sampling in the context of this thesis.

It is not uncommon to have mobile call graphs representing very large networks, composed by millions or even billions of nodes. Performing a complete analysis of these graphs becomes a challenging task, if not unfeasible, due to the additional computational complexity. Considering that vertex count has a time demand of $n$ and edge count of $m$, one might see that the complexity of computing centrality measures is, at least, $O(mn)$, triangle counting is $O(m^{1.41})$, and eigenvector computation is $O(n^3)$, just to mention a few. It may not be possible to compute metrics for large networks in a reasonable amount of time. Moreover, if the aim is to develop a computationally intensive algorithm that is to run over a large graph, testing would be made easier if one could have access to a smaller similar graph where tests are easily ran [30].

In order to overcome these computational obstacles, it is often useful to perform network sampling. As mentioned by Ahmed et al. in [31], “Network sampling is integral to the analysis of social, information, and biological networks. Since many real-world networks are massive in size, continuously evolving, and/or distributed in nature, the network structure is often sampled in order to facilitate study.” As a matter of fact, there are several purposes to network sampling [32].

To begin with, one may want to sample a representative set of vertices or edges, with the objective of estimating the properties of those in the original network, as for example, estimating node degree or degree distribution. Alternatively, there might be the case of wanting to analyze a small subnetwork that is sampled from the original network, while aiming to maintain its structural properties, namely, degree distribution, clustering coefficient distribution, community structure, among others. In this case, and given the reduction of network size arising from the sampling process, it is important to consider scale-down effects on network properties. For example, the average node degree of a sampled network should be smaller than the one of the original network. Finally, local substructure sampling can also be used when estimating their relative frequency or count in the large network. For instance, one might want to understand how many triangles there are in a large graph.

The scenario in which network sampling is to be performed, is also extremely relevant. In some situations, there is full access to the network, meaning that the entire set of nodes and edges is available for random selection. In other cases, only restricted access may be provided. As an example, there are networks in which a seed node or set of seeds is provided and one has only access to its immediate, one-hop neighbors. From there, the network can be explored by accessing the immediate neighbors of a node that is being visited, however, a visualization of the entire network is not allowed. Lastly, some networks have a dynamic structure which changes over time. Such is the case of constructing a graph with data that comes from streaming. In this scenario there is limited memory space and fast moving data, often arriving in an arbitrary order.
A wide variety of sampling techniques has been explored, considering each sampling purpose and scenario. In the context of this thesis, the interest point lies on obtaining a representative sample with similar structural properties as the original graph, assuming full access to the network, where it is feasible to develop and test inference algorithms. Bearing this in mind, there are three essential approaches to graph sampling: node sampling, edge sampling and exploration sampling.

Beginning with node sampling, there is the uniform approach or Random Node (RN), in which nodes are selected uniformly at random [32]. The sample graph will be the induced graph, meaning that it will contain the set of randomly selected nodes and the edges between them. This method is generally biased to low degree nodes and does not present a good match for the degree distribution of the original large graph. Alternatively, non-uniform methods have also been developed. Random Degree Node (RDN) is a variation in which nodes are selected randomly with a probability proportional to its degree, and Random PageRank Node (RPN) will select nodes with a probability proportional to the node’s PageRank value. The last two methods are biased toward high-degree nodes.

Edge sampling can also be performed uniformly, with the Random Edge (RE) algorithm. An edge is uniformly selected at random and the sample graph will also contain all nodes that serve as endpoints to the set of selected edges, in the large graph. If the sample size is small, which it normally is, the outcome is generally a sparcely connected graph, with no respect for the original community structure. Additionally, the sample graph will be biased towards high-degree nodes, since those are the ones with the higher number of incident edges, and, therefore, more likely to be included. To overcome this last drawback, it is possible to use the Random Node Edge (RNE) algorithm, which randomly selects a node and then an edge incident upon it, sequentially.

A different alternative is sampling by exploration, also very commonly used in the context of having restricted access to the large network. The underlying idea is to select a node uniformly at random and then proceed by exploring the nodes in the neighborhood. It entails two different types of methods: random walks and traversal based sampling methods. The difference lies in the format of exploration – the former methods perform exploration with replacement, while the latter do so without replacement. In other words, random walks are memoryless in the sense that nodes can be visited more than once while the algorithm is running, and the same does not happen for graph traversal techniques. The Random Walk (RW) algorithm randomly picks a starting node – also referred to as seed node – and simulates a random walk in the graph. At each hop, it returns to the seed node with probability $p$ (usually 0.15), and restarts the random walk. One clear issue is that the algorithm is susceptible to be stuck in a local dense region if, for example, the starting node is a sink or belongs to a small isolated component of the large graph. In order to surpass this problem, a variation called Random Jump (RJ) performs similarly but instead of returning to the original node with probability $p$, it jumps to another random node of the graph. RW and RJ are both biased towards high-degree nodes and so, to ensure a uniform sampling of nodes, one might try the Metropolis-Hastings Random Walk (MHRW). It works in the same way as RW but introduces the Metropolis-Hastings correction to accept or reject a proposed move in the random walk, thus systematically eliminating the bias. However, “the application scenario of the MHRW is more limited than ordinary RW. In order to calculate the transition probability one has
to know the degree of neighboring vertices”. There exist other variations of the RW algorithm, as seen in [33], but for the sake of this thesis they will not be detailed. As for graph traversal sampling, it is also referred to as topology based sampling. Breadth-First Sampling (BFS) and Depth-First Sampling (DFS) work with the concept of queues. A seed node is initially placed in the queue. At each step, one node is retrieved from the queue, placed in the sample set, and its neighbors that have yet to be visited (and which are still not in the queue) are placed in the queue. The variations regard the way a node is retrieved from the queue: BFS retrieves the first element, while DFS retrieves the last one. BFS discovers all vertices that are at a distance $d$ before discovering any vertices at a distance $d+1$, and DFS discovers the farthest vertex along a chain, then backtracks. Snowball Sampling (SS) is commonly used in sociology studies. It works similarly to BFS: an initial seed or set of seeds is given and at each stage $k$ neighbors of the previously visited nodes are collected. Lastly, there is the Forest Fire (FF) algorithm, which was originally proposed in [34] as a graph generation model. It consists of a probabilistic version of the SS in which $k$, the number of selected neighbors, is geometrically distributed. In $K \sim \text{Geometric}(p)$, if $p$ is set to be $1/k$, then FF becomes SS. All the graph traversal techniques do not allow uniform node selection [35].

Hu et al. mention in [33] the relation between property preservation and property estimation: “[...] the relationship is not symmetric. Namely, property preservation is actually a stronger and more useful notion.”. The point being that, when property preservation is guaranteed, property estimation methods on sampled graphs can lead to good results. On the contrary, “not all property estimation results can be easily casted to property preservation results”. Thus, RN, even though suitable for property estimation, does not perform well in this context, since it generally fails to preserve many properties of the original network (diameter, hop-plot, clustering coefficient, degree distribution and so on). RDN and RPN present the same challenge, whilst introducing a bias towards high-degree nodes. As for ES and RNE, the issue is the sparsity of the sampled graphs, failling to match the original large network. In [35], Leskovec et al. presents the results for the sampling methods performed in five different data sets. Each algorithm (RN, RDN, RPN, RE, RNE, RW, RJ and FF) is run ten times on each of the five real graphs. They conclude that, in terms of degree distribution RE presents the worst fit to the true distribution. RW and RJ, being biased towards high degree nodes, present an over-estimation of the tail of the degree distribution, at the cost of an under-representation of the lower-degree nodes. They further observe that FF is not biased towards high-degree nodes and that the degree distribution is well matched. Regarding clustering coefficient preservation, RE and RNE are not a good fit. Given a small sampled graph, sparse sample graphs produced by these algorithms barely present any triangles. The same is not verified for RW, RJ and FF. In conclusion, they present FF as the best performing algorithm for sampling in large graphs. Sarmento et al. mention SS as one of the most commonly used methods for static network sampling [36], even though this method introduces a slight bias towards the region of the network to which the seed node belongs to. Exploration methods in general, are appointed in [32] as being biased towards high-degree nodes and usually underestimating the clustering coefficient. They are believed to provide a relatively good match of the degree distribution of the original network and the sampled graph can usually preserve
the original diameter, provided that enough nodes are visited.

**Table 3.1:** Sumary of network sampling methods: minimum assumption, sampling objective and properties preserved. FA - full access to network; PA - partial access to network; GC - graph connectivity; PE - property estimation; PP - property preservation; DD - degree distribution; CC - clustering coefficient; CM - centrality measures

<table>
<thead>
<tr>
<th>Sampling method</th>
<th>Assumption</th>
<th>Objective</th>
<th>Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node</td>
<td>FA PA GC PE PP DD CC CM GC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RN</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>RDN</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>RPN</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edge</td>
<td>RE</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>RNE</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exploration</td>
<td>RW</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>RJ</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>BFS/DFS</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>SS</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>FF</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Table 3.1 summarizes the comparison between different network sampling methods. Node sampling techniques have difficulty in presenting network properties, in particular degree distribution, but serve the purpose of property estimation. Similarly, edge sampling techniques also have a hard time preserving network properties, in particular clustering coefficient and graph connectivity. Conversely, exploration sampling methods manage the preservation of degree distribution, even though with a slight bias towards high degree nodes, which can lead to underestimation of the clustering coefficient.

On the whole, there is no perfect sampling method for each particular sampling purpose and scenario. The choice of method should bear in mind the relevant property or properties to preserve. Nonetheless, [32] points out that unbiased sampling is typically not a good choice when one wants to find a smaller graph similar to a large network, with full access to it. It is, then, possible to say that methods such as SS and FF, notwithstanding the non-uniform nature of node selection, provide good fits to such problem. As a final note, there are some remarks to be made concerning the size of the sampled network. In [30], Frank et al. conclude that there is a threshold below which representative graph sampling is not possible and that is around 15%. The study performed by Leskovec et al. also shows that “a 15% sample is usually enough, to match the properties of the real graph”.

In the context of this thesis, the original network’s dimension and its consequent complexity constitute a challenge for the testing of inference algorithms. As such, network sampling is an important initial step. Given that the objective was to find a smaller network that more or less preserved the network properties of the original network, specifically degree distribution and clustering coefficient, the adopted method was Snowball Sampling. The choice was based on the affirmation in [36] that SS is a commonly used method for static network sampling and the fact that this method is able to preserve the degree distribution and clustering coefficient of the original network.
Exploratory data analysis

Contents

4.1 Introduction ....................................................... 30
4.2 Findings ................................................................. 32
4.3 Conclusions ............................................................ 40
Exploratory data analysis is a fundamental process of any data science project in which the quality of the data is assessed. This chapter begins with an introduction to the concept of exploratory data analysis, followed by a section of analysis of the findings obtained for the dataset available for this thesis. The latter entails a description of the dataset, an evaluation of its variables and a section with the presentation of social strategies found through the process. A final conclusion of the analysis is included at the end.

4.1 Introduction

Telecommunications operators have access to their users patterns of communication through the Call Detail Records. As previously stated, these records hold information about communications between two mobile phones: the identification of the intervenient mobile phone numbers, the direction of the communication, the type of communication, the timestamp and duration, and, possibly, the localization of the cell that operated the communication. Such records can be fully anonymized and gathered into datasets in order to permit an analysis of the information they convey. As stated in [3], “Data Science is an exciting discipline that allows turning raw data into understanding, insight, and knowledge”, and that is precisely the scope of this thesis. Leveraging from the Data Science (DS) field, this thesis proposes an analysis of fully anonymized datasets provided by telecommunication operators in order to obtain knowledge regarding the communication patterns of the users and how they relate to their demographics. An inference method is presented so as to use such insights to be able to predict user demographics for a wider set of users. Gaining knowledge through inference on clients that may not have provided their demographic informations to the telecommunication operators, creates the opportunity for a more efficient, personalized service.

A typical DS project has well sectioned stages: one begins by importing the data to be analyzed, which one then has to tidy according to the objective of the analysis, followed by an understanding stage and a final communication of the findings. The whole process can be programed using the R language and tools and a scheme is presented in Figure 4.1 proposed in [3].

![Figure 4.1](image)

**Figure 4.1:** Scheme of the format of a DS project. The DS project will be running on a program which will encompass different stages that go from importing the data, tidying it, understanding the data and communicating the findings. During the understanding stage, it is commonly necessary to transform, visualise and model the data iteratively, so that information can be extracted. *source: [3]*

After importing or gathering the data to be analyzed, it is often necessary to tidy it, since it does not always come in the format that best suits one’s analysis. In this stage one can utilize the tidying
tools and packages provided in R. The idea is to consistently organize the data so that it becomes easier later on to work with it. Possible actions include spreading or gathering variables, separating or uniting variables and dealing with missing values. Particularly, one might want to change the format of a variable, for example, discretizing a continuous variable by binning its values. One commonly used variable format in R is a Factor variable – it means that such variable will only assume one of a pre-defined discrete set of values (factors). Furthermore, it might be useful to not only change the format of the variables but also change the structure of the dataset, for instance, merge two datasets or separating one into more datasets. It is also somewhat common to have missing values in the available data. Missing values can appear in datasets explicitly, when they are directly flagged with ‘NA’, or implicitly, whenever the the value for a variable in specific observations are simply not present in the data. It is important to be careful when performing mathematical computations with variables that might have missing values, since the result will be compromised.

Once the data is tidy, one can begin the understanding phase and try to extract information. This process is also referred to as Exploratory Data Analysis (EDA), where the idea is to follow an informal structure of work in order to investigate the quality of the data: one should generate questions about the data and search for answers through the transformation, visualization and modeling of the data; once some information is learnt, questions can be refined and the process can be repeated iteratively, until some final conclusions are drawn. Transformation actions include filtering specific rows or selecting specific columns, arranging the observations in certain orders, grouping and summarizing or maybe computing new variables from the existing ones. As initial steps of the EDA it might be useful to look into what kind of variations occur within the variables and analyze the covariation between pairs of variables in the data set. With such objectives in mind, visualization tools come in handy: there are different plotting tools available in R to visualize the distribution of variables in the dataset, whether being discrete or continuous variables. There exist also various plotting options to analyze the correlation between those different types of variables.

When performing EDA, one possibility is to encounter imbalances in the dataset. Datasets typically reunite information about a specific reality and the observations registered refer to different elements of that reality. Those elements can be labeled into different classes, explicitly or implicitly, through the variables of the dataset. For example, one may have a dataset with information on bank transactions and may want to distinguish between two classes: legal or fraudulent transactions. The class of each transaction observed will be determined by the variables that describe it in the dataset. As another example, in a dataset describing users of mobile phones, users can be labeled as male or female according to a single variable gender. Having an imbalanced dataset means that there is an overrepresentation of some classes in detriment of others. More often than not, real data comes in the form of imbalanced datasets. It is certain that many different factors can affect the balance of a dataset and, consequently, constitute sources of biasing, making its analysis all the more challenging. Imbalanced datasets are usually problematic for performing learning tasks, as is the example of classification of transactions into legal or fraudulent. Nonetheless, there are techniques that can be used to try to overcome this challenge or at least, lessen its consequences to inference,
as pointed out on [37]. Options include minority over-sampling or majority under-sampling methods, using penalized classification models in which a cost is attributed to samples of different classes or using already developed algorithms for learning from skewed data. Taking fraud detection problems as the example, datasets are most likely imbalanced and there are studies such as [38] and [39] that address this problem.

EDA is frequently performed in many DS projects. It is extremely relevant to be able to understand and assess the quality of data prior to modeling it for any inference purpose.

4.2 Findings

4.2.1 Description of the mobile call dataset

Two different, but complementary, fully anonymized datasets were provided for this thesis by a portuguese telecommunications operator. The first dataset amounts to the communication records and the variables are as follows: anonymous identifier of the origin phone number, anonymous identifier of the destination phone number, total duration of calls made, total number of calls made and total number of sms sent from origin to destination. Absolutely no communication content was disclosed anywhere in the dataset. Additionally, there is another dataset with demographic information which was collected for some of the operator’s clients. It is composed by anonymous identifier of the phone number, anonymous number of client associated to the phone number, age and gender of the client. Since the mobile phone user’s numbers are encrypted, identities are unrecoverable. Furthermore, the gender and age labels were obtained by the telecommunications operator with a disclaimer of being noisy. Due to privacy contraints, the sources of such labels were never questioned in the context of this thesis. The data was collected in 2013 over a period of 6 months.

After pre-processing – tidying the variables – and merging both datasets, the result was a mobile call dataset with the crossing between anonymous identifiers of phone numbers in the communication records and the associated demographic information in the additional dataset. Naturally, some of the intervenients of the communications registered were left unlabeled, since the demographic information associated to such phone numbers’ anonymous identifiers was not available – this can happen either because they didn’t provide the information or because they are clients of other operators, while still communicating with clients of this particular operator and, thus, appearing on the CDRs. Both age and gender are factor variables, meaning they assume one of a discrete number of values. Gender is either male or female, while age is divided into 4 age groups, defined in accordance to the business needs of the operator: young (under 25), young adult (aged 25 to 35), middle aged (aged 35 to 55) and senior (over 55 years old). As for the remaining variables, the number of calls and the number of sms are integers as well as the total duration of calls, representing time in minutes.

The complete mobile dataset has around 3.9 million observations of communications between pairs of origin and destination. In the light of reducing the complexity of algorithm testing during this DS project, a smaller sample, as mentioned in chapter 3, was used both for EDA as well as for inference purposes. The sample mobile call dataset is composed of approximately 157.8 thousand
distinct communication records, and around 50% of the observations had missing information, namely labels for the age and gender variables, on one of the intervenientes. From there, a mobile call graph was built considering at least one bidirectional communication – either sms or calls over 5 seconds – between each pair of mobile phone users and only the largest connected component was examined. The resulting network is composed by 17334 nodes and 18724 edges connecting them.

As stated before, the objective of this thesis is to use behavioral patterns and social strategies found in mobile call graphs as a stepping stone for predicting the age and gender of mobile phone users. With that in mind, it is crucial to analyze the dataset available in order to know if it is, in fact, possible to identify some of those behavioral patterns that might help distinguish between different demographic groups: males, females, young, young adults, middled aged and senior users.

### 4.2.2 Variable Analysis

When performing EDA, a natural initial step is to look into the dataset and individually assess which trends and observations can be retrieved for each variable. As can be observed in Figure 4.2, both gender and age distributions have a critical impact on the work, since they present some unexpected characteristics. Beginning by gender (Figure 4.2a), there is a strong presence of male mobile phone users. Approximately 66% of the labeled clients, i.e. disregarding missing values, are identified as male. Such might be explained by a bigger affluence of male clients into this telecommunications operator’s business, or it may arise from the fact that some client accounts can hold more than one phone number, in which case the demographics of the client that signed the contract will be associated to all the phone numbers used by the remaining members of the aggregate. Regardless of these assumptions being the reason for imbalance or not, which cannot be confirmed in this EDA, or even if there existed other explanations for the gender imbalance in the dataset, there is a clear bias introduced in the study of this mobile call network. Age distribution, again, presents another source of bias. As depicted in Figure 4.2b, there is an overrepresentation of the older age groups when compared to younger generations. It might, once again, arise from different business characteristics.

Figure 4.2: Distribution of the gender (a) and age (b) variables in the mobile call dataset. There is an overrepresentation of male over female and of older generations over younger generations of mobile phone users in the dataset, which will introduce an important bias factor in the data analysis.
of the telecommunications operator. In fact, cross-referencing the demographic distributions with the
tones registered for the entire population of the country and at the time the CDRs were collected, they
do not match. Gender distribution was relatively uniform and, in fact, there were more females than
males [40]. There was not such a great presence of people over 35 years old, whilst there was a
bigger presence of young people [41]. That being said, such biasing factors need to be taken into
account when performing the inference study.

Figure 4.3: Mean values of the variables number of calls (a) and number of sms (b) for each demographic group.
It becomes evident that calling is the preferred communication method over exchanging sms and that females
tend to communicate more often than men. Additionally, younger generations have a higher predisposition to
use sms, even though they receive more communications than the ones they make, when compared to older
generations, which prefer calling.

Regarding the communication variables, Figure 4.3 shows the mean number of calls and sms,
incoming and outgoing, specified for the different demographic groups. A tendency to communicate
more often via call than via sms is clearly observed, which may be justified by the increased use of
mobile internet applications with chat capacities. In particular, the fact that the average number of
sms sent or received by a user in this dataset is smaller than 1 tells a great deal about the utilization
of this service, which is slowly decreasing [42]. Both Figure 4.3a and 4.3b show that females typically
communicate more often than males. As far as age trends go, older generations prefer communicating
through calls, when compared to younger ones. However, even though younger generations have a
higher predisposition for communicating via sms, they are more likely to receive an sms than to send
one. This might, again, be explained by a more frequent use of internet applications. In terms
of incoming and outgoing communications they are typically well balanced, except for the case of
number of sms sent/received by younger generations. It is very important to bear in mind that the
sample of users labeled as young and young adult is significantly smaller than the others, which
naturally impacts the fairness of the results shown.

From the mobile call dataset it is also possible to compute the average duration per call, dividing
the total duration by the number of calls between each pair of users and finding the mean value. On
average, people tend to make calls with a mean duration of 142.4 minutes. In Figure 4.4 it is possible
to distinguish a tendency for older generations to participate in longer calls when compared to the
younger groups, but this is a variable which is relatively well balanced over the different demographic
groups.
Average duration of a call calculated by gender and age. Even though the average duration of a call is well balanced throughout all demographic groups, there is a small tendency for older generations to participate in longer calls, when compared to younger ones.

4.2.3 Social Strategies

Similarly to what was done in most state of the art studies, the EDA process done over the mobile call dataset allows for a recognition of certain social patterns. Before going any further, it is critical at this stage to remember the possible consequences of the imbalance of the dataset. Since there is an overrepresentation of male over female users and of older over younger age groups, most social strategy analyses will be compromised. However, and emphasizing the fact that real datasets are often imbalanced and not according to one's expectations, it still makes sense to try to identify some of the most commonly identified social patterns pinpointed by the state of the art in the available data.

A most prominent social strategy consensually found in all state of the art analysis is the presence of both age and gender homophily [14], [15], [16], [2]. Homophily is a theory of sociology defined as the tendency that humans have to relate to others with similar characteristics. In the concrete case of mobile call graphs, it is common to find that people tend to contact others of the same gender and age group as themselves. In terms of gender homophily, it can be perceived in the dataset for male

(a)
(b)

Figure 4.5: Distribution of gender of a user’s contacts. Homophily has been confirmed in mobile call graphs, which means females would have more female contacts, while males would have more male contacts. Since the dataset is imbalanced, gender homophily can only be discerned for male contacts (b): the fact that there are mostly males in this network makes it so that women also have more male contacts. Even so, the percentage of female friends in considerably higher for female users (a), which could indicate the existence of general gender homophily in the mobile call graph.
users, yet it is more unclear for female users. In a dataset with predominantly male users, it is natural that females have more male friends than female ones, as seen in Figure 4.5. Still, the percentage of female contacts is higher in the case of female users (4.5a) than of male users (4.5b), which indicates that women tend to connect to other women more than men do. Age homophily is also commonly rec-

![Figure 4.6: Distribution of age of a user's contacts.](image)

ognized and will, once more, suffer the biasing effects of the imbalance of the dataset. The expected trend is observed for older generations, but is not explicit for the young underrepresented generations, as shown in Figure 4.6. That is mostly due to the fact that the young and young adult users correspond to around 10% of the users labeled as middle aged or senior. One way of overcoming this bias and trying to discern age homophily in Figure 4.6 is looking at the percentage of each age group, across all graphics. In fact, one might notice that the percentage of contacts of a given age is bigger for users of that same age. Young contacts are in higher percentage in (a) in detriment of (b), (c) or (d), young adult contacts are in higher percentage in (b) in detriment of (a), (c) or (d) and so on. Overlooking the biasing effects, it is possible to discern small evidences of homophily, both in age and gender, in this

36
mobile call graph.

Another interesting analysis has to do with cross-gender communication behaviors. Relative frequency is computed by dividing the total number of communications existent in the dataset for each gender pair by the total number of those gender pairs. In Figure 4.7 it is possible to see that females communicate more frequently to other females, an average of 11.9 times in 6 months, than males do with other males, who communicate an average of 8.89 times. Interestingly, mobile communications between users of different gender are also very frequent, corresponding to an average of 9.56 times in 6 months. Figure 4.7 highlights not only gender homophily in mobile call graphs, but also cross-gender social patterns, possibly associated to conjugal relationships.

Communications between mobile phone users of different gender is also significant, possibly denoting social relationships of conjugal character.

Figure 4.7: Relative frequency of communication among gender pairs. It stands out that females usually engage in more communications with other females than males do with other males. Notably, the frequency of communications between mobile phone users of different gender is also significant, possibly denoting social relationships of conjugal character.

With regard to cross-generation strategies, they are also visible through mobile call graphs. It is remarkable to notice that, as age increases, mobile phone users tend to communicate less frequently with their similarly aged peers. Figure 4.8 shows that users labeled as young engage in more frequent communications with their similar aged peers. While older generations communicate less frequently among themselves, there is an interesting tendency for young adults to communicate more frequently with seniors, when compared to the remaining cross-generation combinations. There is a high probability that users in those age gaps share relationships of parental character, which is likely to justify higher communication frequencies.

Figure 4.8: Relative frequency of communication among age pairs. It comes as no surprise that users labeled as young engage in more frequent communications with their similar aged peers. While older generations communicate less frequently among themselves, there is an interesting tendency for young adults to communicate more frequently with seniors, when compared to the remaining cross-generation combinations. There is a high probability that users in those age gaps share relationships of parental character, which is likely to justify higher communication frequencies.
between young users and their parents, likely labeled as middle aged or senior, may not be very frequent, as youngster typically communicate with same age peers. However, for pairs of users in age groups young adult and senior, the frequency of communications increases. These age groups would include users related through parenthood, which is one way of explaining an average of 10.79 communications in 6 months.

Alternatively to looking into pair relationships along the mobile call graph, one can also look into social circles. In particular, triads are local graph structures often present in social networks and which consist of subgraphs of 3 interconnected nodes. As mentioned in [43], “Evidence from many sources shows that triadic tendencies are important structural features of social networks [...]”, and as such, can convey important social behaviors. Triadic closure, measured through the clustering coefficient, is a concept of graph theory that suggests that if two nodes A and B are connected and B and C are also connected in a graph, then A and C are likely to be connected as well. It is a property that hardly holds truth globally in very large, real social networks [44]. However, it is also possible to compute local clustering coefficients, and assess to which degree do nodes in a graph tend to cluster together. Apart from computing local clustering coefficients for each graph node, one can additionally study the demographics of nodes included in triads in the mobile call graphs. As observed in Figure 4.9:

![Percentage of labelled triads in which females are included](image)

(a)

![Percentage of labelled triads in which males are included](image)

(b)

**Figure 4.9:** Percentage of users in different labeled triads in the mobile call graph by gender. (a) shows that of the females who are included in labeled social triads in the mobile call graph, most of them are connected in a triad with another female and a male. Contrarily, males are more repeatedly included in triads with all male members (b).

4.9a: females are most likely to engage in groups of two females and a male, while males tend to stick to groups of males only, as seen in Figure 4.9b. Furthermore, Figure 4.10a illustrates the age distribution of females that participate in triads with different configurations, while Figure 4.10b does the exact same for males. With respect to age, the vast majority of users connected in triads is of older generations, suggesting that middle aged and senior users tend to have closer relationships with their contacts. Furthermore, it would be relevant to analyze the triadic behaviors of people in active dating periods, namely young adults. Even though they are underrepresented, one can notice that young adults tend to be included in triads with others of different genders, supporting the assumption that that is the age at which one tends to look for a dating partner. The exception of most young adult
males being included in only male triads, as seen in Figure 4.10b, is most likely affected, yet again, by the imbalance in the dataset.

In [2], Dong et al. also found another interesting social strategy from mobile call graphs. By analyzing network metrics such as degree, neighbor connectivity (the average neighbor degree), clustering coefficient and embeddedness, they concluded that “younger people are active in broadening their social connections, while older people have the tendency to maintain smaller but more closed connections”. Through conducting the same analysis, the finding could not be supported in this dataset. Indeed, as seen in Figure 4.11, degree and neighbor average degree were not found to be higher for younger generations, while clustering coefficient and embeddedness measures where also not significantly higher for older generations, as opposed to what was found in the mentioned state of the art study. In fact, these conclusions could arise from an early adoption of Internet communication application by younger generations, which can substitute the communication functionalities of mobile phones. For this thesis, information on the usage of mobile internet data is not accessible. These observations are also subject to the age variable binning used in this thesis.

In conclusion, several social strategies could be perceived from the mobile call network. Evidence was found for the existence of both gender and age homophily: mobile phone users tend to communicate with others with similar demographic properties. With respect to communication frequency, females communicate more often to other females and young users are the most active of all age groups. Cross-gender communications are also frequent, denoting possible conjugal relationships. Furthermore, relations of parental character could also be translated in a high frequency of communications between young adults and senior users, when compared to other cross-generation communications, excluding those among youngsters. It was also found that, when forming groups of three connected users, males tend to connect to other males, while females are relatively more likely to take part in different gender triads. Additionally, seniors have a higher presence in social triadic circles, when compared to younger generations, implying that, while the latter are more active in their communications, the former engage in closer relationships with their contacts.
Figure 4.11: Network metrics of mobile phone users by gender and age. As opposed to what was found in [2], in this mobile call graph, younger generations have smaller degree (a), than older generations, which suggest the latter have a higher tendency to have more mobile phone contacts. The neighbor average degree assumes a completely different pattern for males and females, possibly affected by the higher percentage of males users in the network. As for clustering coefficient, (c) shows slightly higher values for younger generations and embeddedness (d) is relatively uniform for all demographic groups. Nothing concrete can be concluded about the tendency of having closer relationships for each gender and age group. The network metrics were calculated for a mobile call graph constructed from the mobile call dataset as defined in chapter ? Model and Inference.

4.3 Conclusions

There are two major take aways from the EDA performed on the mobile call dataset. The first one has to do with the imbalance of the dataset and the impact it will have on the work, and the second one relates to the social strategies that can, nonetheless, be discerned in the mobile call graph. As in most real cases, data rarely comes exactly as one expects. Even when not aiming for specific conclusions to be drawn from data, one often encounters data that is hard to work with, for example, having a lot of missing values or demanding extensive pre-processing. This work is the perfect such example: the mobile call dataset was constructed with anonymized information shared by a telecommunications operator and, therefore, corresponds to a supposedly non-fabricated reality. The consequence is a truly imbalanced dataset that questionably represents the reality of the mobile phone users. It was clear that there is an overrepresentation of males over females and of older generations over younger ones, confirming the assertion that the labels provided by the telecommunications operator are noisy. It is also insurmountable that this imbalance has a significant impact in both the analysis of the data and
the following inference problem resolution. That being said, one can only work with what is available. All in all, the purpose of the EDA was not only to assess the quality of the available data, but also to identify to which degree do demographics influence mobile phone user interactions. In fact, evidence was found for the existence of varied social strategies in the mobile call network, which leads to the assumption that the gender and age of mobile phone users are in some way related to their mobile phone usage behaviors within the network. It is based on this premise that a model for inferring mobile phone user demographics from mobile call graphs should leverage from the structure of the network.
5
Experiments

Contents

5.1 Description of the problem ............................................. 44
5.2 Tree-based classification .............................................. 45
5.3 Experiment Setup ..................................................... 47
5.4 Results ............................................................... 48
5.5 Related work .......................................................... 54
In this chapter the problem of inferring user demographics from mobile call graphs is described. Three tree-based models are presented as candidates for experiments. With the goal of evaluating the contribution of network metrics to the prediction of gender and age based on mobile call graphs, these three methods are compared for two scenarios: one which the set of features included network related metrics and one where it does not. Results are discussed and some related work is mentioned in the end.

5.1 Description of the problem

Following the EDA, one can proceed with the inference goal. The idea is to leverage the information contained in the mobile call dataset – used to construct the mobile call graph – in order to predict the demographic class of a user. In other words, the expectation is to be able to determine the most probable demographic class of a particular user, given a set of features extracted for said user and considering the connections with other users in the network. The underlying assumption is that the demographic properties of a mobile phone user, namely age and gender, are related not only with the characteristics of that user (represented in a set of features) but also related to the demographics of the users’ contacts in the network. In fact, and as presented in the previous Chapter 4, the distributions of the mobile call dataset’s variables expose specific communication patterns associated to the users’ demographics, as for example, the fact that females tend to communicate more often than males or that younger generations have a higher predisposition to use sms than any other generation. These findings suggest that mobile phone users’ demographics can be perceived from their communication behaviors, which can be translated in a set of user features. Furthermore, varied social strategies can be conveyed in mobile call networks, such as gender and age homophily or the demographic distribution of social triads, implying that users demographics influence the way they connect with other mobile phone users. That being said, eight different demographic classes can be assigned to users: 1) Male - Young, 2) Male - Young Adult, 3) Male - Middle Aged, 4) Male - Senior, 5) Female - Young, 6) Female - Young Adult, 7) Female - Middle Aged, 8) Female - Senior.

For the purpose of testing inference algorithms, the sampled dataset was used. The mobile call graph is obtained as follows: an edge connecting two nodes, representing users \(u_i\) and \(u_j\), exists in the mobile call graph \(G\) if there is at least one bidirectional communication – either sms or calls with duration higher than 5 seconds – in the mobile call dataset, within the period of observation. Only the largest connected component is considered and the final result is a graph \(G\) with 17334 thousands nodes and 18724 edges. Furthermore, a set of 39 features is extracted for each user present in the network \(G\). Such features are grouped into communication related features, topological features, friend related features and triad related features. Communication features are the ones obtained directly from the mobile call dataset and include number of calls, number of sms and total duration of calls, distinguishing between incoming and outgoing, as well as a computed average call duration. They also include three entropy values denoting the total number of contacts (all, just call or just sms) a user has over the relative frequency at which he/she communicates with them, making up for a total
of 10 communication related features per user. As reported by Montjoye et al. “The more one interacts equally often with a large number of contacts the higher the entropy will be” [7]. Taking advantage of the mobile call graph structure, topological features stemming from graph theory are also considered, including node degree, average neighbor degree, weighted average neighbor degree (weighted by number of communications), clustering coefficient, closeness centrality, betweenness centrality, eigen centrality and embeddedness, in a total of 8 features. Additionally, there are friend related features which consist of 7 variables indicating the number of friends a user has which are labeled with each category (male, female, young, young adult, middle aged, senior) plus a total number of friends that are labeled. Finally, triads are also considered as a relevant network structure based on state of the art and the conclusions of the EDA. As such, the set of features further comprises a total number of triads in which the user is inserted, as well as the number of triads including the user (denoted by v) and two other nodes labeled in any of the possible category combinations: FF-v, MF-v, MM-v, AA-v, AB-v, AC-v, AD-v, BB-v, BC-v, BD-v, CC-v, CD-v, DD-v, where A, B, C and D stand for young, young adult, middle aged and senior, respectively. On the whole, the set of 39 features is aimed at spanning all possible information about each user that can be retrieved from the mobile call dataset and mobile call graph and is based on what was used in the state of the art and considering all findings of the EDA.

Indeed, this problem can be framed as a ML problem of classification: the objective is to classify nodes with a label – or class, as shall be denoted henceforth – among the eight possible combinations of gender and age, given the network structure and the set of features extracted from the dataset for each user. In order to train the algorithm to learn how to classify a user, a portion of the dataset is used as training set. This means that for a percentage of the whole set of users the demographic class will be provided, while the remaining percentage will be used as test set to analyze the prediction capacities of the algorithm. An accuracy will be reported based on such predictions.

5.2 Tree-based classification

In order to evaluate the impact of network metrics in the prediction of mobile phone user’s from call graphs, some classical ML classification algorithms were implemented. The problem consists of predicting a class for each user, within the discrete set of possible demographic classes, considering the user features extracted from the data. In that sense, classification algorithms based on classification trees are suitable to this analysis.

Decision trees are predictive models that rely on a tree structure in which branches represent conjunctions of features that lead to the class labels, represented by the leaves. Essentially, the process of selecting a class for a given instance on the dataset can be described as sequence of binary decisions corresponding to the traversal structure of a tree. Tree-based methods work by segmenting the prediction space into a simpler set of regions and the rules that define the segmentation can be clearly expressed in its structure, as illustrated in Figure [5.1]. These methods are regarded in [13] as combination methods in which a single model is responsible for making the predictions at a given point.
is the prediction space. Such single models are selected, given an input, through a process that is described by sequential decision making. In order to learn the models responsible for prediction each target variable from a training set, it is necessary to determine the structure of the tree. Particularly, the predictor variable that defines the split decision at each node of the tree and the threshold values $\theta_i$ must be defined. “Even for a fixed number of nodes in the tree, the problem of determining the optimal structure (including choice of input variable for each split as well as the corresponding thresholds) [...] is usually computationally infeasible due to the combinatorially large number of possible solutions” [13]. The common approach is to use a greedy optimization by starting with a root node, corresponding to the whole prediction space and then growing the tree, adding one node at a time. At each step, a number of regions will be candidate to be split, corresponding to the addition of a pair of leaf nodes to the existing tree. Depending on whether one wants to find a quantitative or a qualitative prediction, one can use regression trees or classification trees. Either way, trees have a significant advantage: they are easily interpreted, even by non-experts. In addition “some people believe that decision trees more closely mirror human decision-making than do other regression and classification approaches” [45]. Therefore it makes sense to initially approach this classification problem with a classification tree method. Dataset features are used as predictors that defined the prediction regions and each observation will be classified with the most commonly occurring class of the training observations that fall in the same prediction region. Unfortunately, trees usually suffer from high variance, which means they are likely to overfit the training set and consequently yield low levels of predictive accuracy. “However, by aggregating many decision trees, using methods like bagging, random forests, and boosting, the predictive performance of trees can be substantially improved” [45].

Random Forests (RF) are a classification tree-based method using Bootstrap Aggregation (Bagging) that significantly improves the predictive results over Decision trees. Bagging is a procedure used in statistical learning methods based on the conception that averaging a set of observations reduces the variance of the model. Hence, the goal is to take many training sets from the data, build separate prediction models for each one and average the prediction results. Since it is not common to have more than one training set available, bootstraping is an alternative that consists of taking repeated samples from a single training set. In the context of classification trees, instead of using a single decision tree, RF construct several decision trees at training time, assuming as prediction the
mode of the classes outputted by the individual trees. In particular, the trees are built in a way that decorrelates them from each other: “each time a split in a tree is considered, a random sample of \( m \) predictors is chosen as split candidates from the full set of \( p \) predictors” [45]. This will make sure that strong predictors do not get used in the top split in most trees, causing them to be very similar to each other, which would bring small added value. Since using many different decision trees, these methods improve the predictive predictive accuracy at the expense of interpretability.

Another interesting approach to decision tree-based methods is that of Gradient Boosted Trees (GBT). Contrarily to RF, GBT grows trees sequentially instead of in parallel, making it so that each tree is grown using information from the previous tree. Boosting is a general-purpose technique, similar to Bagging, that allows combining multiple base classifiers to improve the performance of each single classifier [13]. Furthermore, boosting does not rely on bootstrap sampling, but rather fits each tree to a modified version of the original dataset. Boosting classification trees perform in a complex way, for which the details are omitted in this thesis. One difference between boosting and RF, however, is that “in boosting, because the growth of a particular tree takes into account the other trees that have already been grown, smaller trees are typically sufficient” [45]. Stochastic gradient boosting was initially proposed by Friedman in [46] and served as motivation for the application to classification trees.

5.3 Experiment Setup

Decision trees, RF and GBT models were implemented for gender and age prediction separately. The rational is that gender prediction, being binary, should be easier to implement and understand. From there, multi-class classification is performed to predict which of the four age labels should be attributed to each user. Regarding the methods elected, Decision trees was implemented as a starting point method for interpretability purposes, whilst RF and GBT are state of the art approaches for classification problems.

In both prediction cases the imbalance of the dataset was an issue to be tackled. Balancing of the dataset was done using two distinct techniques: up-sampling and Synthetic Minority Over-sampling Technique (SMOTE). Up-sampling is a technique for balancing datasets that randomly replicates instances in the minority class. SMOTE, however, is a technique for constructing classifier models from imbalanced datasets proposed by Chawla et al. in [47], in which the over-sampling of the minority class is performed by creating "synthetic" examples, as opposed to over-sampling by replacement. In this way, the balancing of the dataset is done “by operating in “features space” rather than “data space”.” The results obtained for the experiments in [47] showed that the SMOTE technique can, in fact, improve the accuracy of classifiers for a minority class. Here, the results are only presented using SMOTE for gender prediction, given that it slightly out-performed up-sampling, and using the latter for age prediction.

As previously mentioned, the anonymized dataset used for experiments comprises 17334 unidentifiable mobile phone users, but the experiment was only conducted for the 15419 users that were
labeled with gender and age. The 40 different variables used in these classification methods are gender or age as a label and the set of 39 features extracted for each user, as disclosed in Section 5.1.

A small change to the feature set was made: since only some friends were labeled, it was thought to be more informative to use percentage of labeled friends of a certain label rather the actual absolute number. That decision was made hoping that the model could better differentiate examples such as user $u$ having one female friend who is the only labeled friend out of five, and the gender of the other friends is unknown; user $i$ having one female friend, and all the other four labeled friends are male; or user $j$, having one friend in total, and that friend is labeled as female. The same line of thought applies to age prediction. Furthermore, 4 fold cross validation was used – where 75% of the data was used for training and the remaining 25% for testing, and the process was repeated four times – and 10 fold cross validation – using 90% of the data for training and 10% for testing and repeating the process ten times. Even though SMOTE was used to balance the training set, the test data was left imbalanced to mirror the actual problem. Confusion matrices (Figure 5.2) are commonly used in statistical learning as a performance evaluation metric. From the numbers presented in confusion matrices it possible to compute several other metrics: accuracy is given by the number of all correctly predicted instances divided by the total number of instances on the dataset (5.1); sensitivity, also known as recall, is the number of correctly predicted positives, divided by the actual number of positives (5.2); and specificity is the number of correctly predicted negatives over the total number of negative instances on the dataset (5.3).

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{TP + TN}{P + N} \quad (5.1)$$

$$\text{SN} = \frac{TP}{TP + FN} = \frac{TP}{P} \quad (5.2)$$

$$\text{SP} = \frac{TN}{TN + FP} = \frac{TN}{N} \quad (5.3)$$

5.4 Results

Following the discoveries of the EDA in Chapter 4, in which social strategies were identified based on the structure of the network and the way users interact within it, the goal is to further evaluate the improvement on inference performance of the classifiers by considering network metrics as predictive
features. As described in Section 5.1, a set of 39 features was defined for each user on the network. Apart from basic communication features, a set of topological features was used with the idea of leveraging from the mobile call graphs’ structure and its properties, as hinted by the results of the EDA. Moreover, triad related features are aimed at translating the demographic relations observed for users that are inserted in this particular graph substructure that are triads (also known as triangles). The hope is that network metrics such as the ones mentioned will improve predictive results even of simple methods such as classifiers.

Beginning with gender prediction, of the labeled instances, 5237 were labeled as female and 10182 were labeled as males before balancing. Females were treated as class 1 (positive), while males were treated as class 0 (negatives). The confusion matrices are presented for the three

![Confusion matrices for gender prediction with the three models using 4 fold cross validation.](image)

**Figure 5.3:** Confusion matrices for gender prediction with the three models using 4 fold cross validation. Considering the imbalance of the test set, it would be desirable that the algorithms predicted correctly a higher number of males (TN) than females (TP), which happened for both Decision trees and RF, but not for GBT.

![Confusion matrices for gender prediction with the three models using 10 fold cross validation.](image)

**Figure 5.4:** Confusion matrices for gender prediction with the three models using 10 fold cross validation. As in the case of 4 fold, both Decision trees and RF predicted correctly a higher number of males (TN) than females (TP), while GBT presented similar numbers.

**Table 5.1:** Sensitivity, specificity and accuracy values for gender prediction with the three models, using 4 fold and 10 fold cross validation.

<table>
<thead>
<tr>
<th></th>
<th>4 fold cross validation</th>
<th>10 fold cross validation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sensitivity</td>
<td>Specificity</td>
</tr>
<tr>
<td>Decision trees</td>
<td>0.35802</td>
<td>0.73415</td>
</tr>
<tr>
<td>Gradient Boosted Trees</td>
<td>0.70746</td>
<td>0.37782</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.45713</td>
<td>0.67119</td>
</tr>
</tbody>
</table>

Methods with 4 fold cross validation in Figure 5.3 and 10 fold cross validation in Figure 5.4. Sensitivity, specificity and accuracy are reported for each method in the 4 fold and 10 fold cross validation tests of gender prediction in Table 5.1. Naturally, 10 fold cross validation allows for slight improvements.
on accuracy, since a bigger portion of the data is being used for training the algorithms. The best performing algorithm for gender prediction with Decision trees, curiously the simplest of all three methods. One might say that more complex inference techniques would benefit from equally more complex features. The variable importance of each predictor is shown in Figures 5.5, 5.6 and 5.7.

![Variable importance for gender prediction using Decision trees](image)

**Figure 5.5:** Variable importance for gender prediction with Decision trees method using 4 fold and 10 fold cross validation. Network related features such as entropy of sms contacts, total number of triads and clustering coefficient appear among the five most important prediction variables.

![Variable importance for gender prediction using GBT](image)

**Figure 5.6:** Variable importance for gender prediction with GBT method using 4 fold and 10 fold cross validation. In the case of GBT, eigen centrality and closeness centrality appear among the more important variables, but are outnumbered by communication features.

When using multiple trees in BGT and RF, variable importance can be calculated through the Gini index \[45\]. Indeed, network metrics such as clustering coefficient, eigen and betweenness centrality, and entropy of sms contacts are among the most important variables across the three methods. However, communication features are also assumed by these classifiers as good predictors. The number of incoming sms is consistently the best predictor for Decision trees and GBT. As mentioned on the EDA, sms technology is becoming less utilized and so most users send and receive zero sms in the observation period of this dataset. Consequently, if a user did send or received sms, he/she become easily differentiate from the rest of the users and thus, this feature is clearly identifiable by
For RF, eigen centrality and closeness centrality are among the most important predictive variables.

As for age prediction, of the labeled instances, 159 were labeled as young, 1244 were labeled as young adult, 7616 were labeled as middle aged and 6400 were labeled as senior before balancing.

Each label young, young adult, middle aged and senior corresponds to class 1, 2, 3 and 4, respectively. The confusion matrices are presented for the three methods with 4 fold cross validation in Figure 5.8 and 10 fold cross validation in Figure 5.9. Sensitivity is computed for each class similarly to (5.2) as the number of instances correctly labeled with each class divided by the actual number of instances that exist of that class. Table 5.2 reports the values of the sensitivities and accuracy for 4

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>class 1</td>
<td>class 2</td>
<td>class 3</td>
<td>class 4</td>
</tr>
<tr>
<td>Decision trees</td>
<td>0.06314</td>
<td>0.31029</td>
<td>0.40467</td>
<td>0.43906</td>
</tr>
<tr>
<td>Gradient Boosted Trees</td>
<td>0.02500</td>
<td>0.05868</td>
<td>0.61489</td>
<td>0.52844</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.02516</td>
<td>0.03617</td>
<td>0.65756</td>
<td>0.53969</td>
</tr>
</tbody>
</table>

Figure 5.7: Variable importance for gender prediction with RF method using 4 fold and 10 fold cross validation. For RF, eigen centrality and closeness centrality are among the most important predictive variables.

Figure 5.8: Confusion matrices for age prediction with the three models using 4 fold cross validation. Consistently the algorithms predicted correctly more instances of classes 3 and 4, which are the majority classes.
Table 5.3: Sensitivity and accuracy values for age prediction with the three models, using 10 fold cross validation.

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>class 1</td>
<td>class 2</td>
</tr>
<tr>
<td>Decision trees</td>
<td>0.06250</td>
<td>0.27334</td>
</tr>
<tr>
<td>Gradient Boosted Trees</td>
<td>0.00625</td>
<td>0.04828</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.01250</td>
<td>0.03137</td>
</tr>
</tbody>
</table>

fold cross validation and Table 5.3 for 10 fold cross validation tests. Again, 10 fold cross validation allows for slight improvements on accuracy and RF is the algorithm that yields the best performance for age prediction. The variable importance of each predictor is shown in Figures 5.10 and 5.11 and

5.12 In this case, the most powerful predictive variable for each algorithm differs among eigen centrality, total duration of outgoing calls and average total duration of call. Closeness centrality is also considered among the most important variables for all three methods. Clearly, eigen centrality and closeness centrality are relevant features for age prediction using tree-based classifiers. Yet, there are still a lot of communication features regarded as best predictors.

It is curious to see how sms behavior seems to have a greater impact on gender prediction, while call behavior allows to better discern between age classes. In the end, Decision trees and RF presented the best performances for gender and age prediction, achieving accuracies of 62% and 55%,
respectively. The dataset is imbalanced, as discussed on Chapter 4, and there are very few instances of user labeled as young or young adult, which leads to all classifiers failing to identify these minority classes, as translated by the very low values of sensitivity for classes 1 and 2. Tree-based methods are clearly suitable for this predictive analysis, but still leave room for performance improvement. Even though the results show that for both age and gender classification using tree-based methods, network metrics such as centrality measures, entropy values or clustering coefficient can be regarded as good predictors, there is not a clear differentiation of communication and network related variables importance. That being said, and in order to better assess the relevance of these added features in the predictive analysis of call graphs, the same tests were repeated, only this time the set of features considered was different. Topological features, including clustering coefficient, closeness centrality, betweenness centrality, eigen centrality and embeddedness; the entropy features, for sms and calls, calls only and sms only; and triad related features were not included, leaving a total of 18 variables (17 features plus the gender or age label). A performance comparison is presented for all three classification methods when considering and when disregarding the network related features. The results are only shown for 10 fold cross validation experiments since they yield slightly best performances than 4 folds. For gender prediction (see Figure 5.13) there is a marginal improvement on accuracy when considering network metrics, but in the case of age prediction (see Figure 5.14) it is not significant. It is then possible to conclude that tree-based classifiers with simple features, such as the ones tested, do not have the capability of taking advantage of the social network structure of call graphs, whose importance is not only confirmed by state of the art, but was also supported by the exploratory data analysis performed on this dataset. The results of these experiments can be explained by different assertions. To begin with, due to the imbalance of the dataset, both in terms of gender and age labels, the insufficient examples of the minority classes, particularly for age, do not permit inference results with significant quality. Secondly, the dataset provided for this thesis has, as disclaimed by the provider, noisy labels. Consequently, there is no guarantee that the features extracted for each user are, in fact, always associated to the gender and age labels attributed to such

Figure 5.11: Variable importance for age prediction with GBT method using 4 fold and 10 fold cross validation. Like in the case of Decision trees, eigen centrality and closeness centrality are among the most important variables.
user. This presents a clear obstacle for simple classification methods to properly learn from the training set provided. More complex features might possibly improve the accuracy of classifiers, however that is yet to be confirmed. In addition, other state-of-the-art studies had access to datasets in which information for the timing of the communications was also provided, making it possible to create separate networks for on and off-hours. Maybe that additional property of the mobile communications could be more informative for the context of network features, since it is known that social behaviors change from day to night time.

In this line of thought, the objective of this thesis is to also provide a baseline for the development of smarter modeling approaches to the problem of inferring user demographics from mobile call graphs. In particular, probabilistic graphical models (PGM) provide powerful inference tools that fully leverage the graphical structure of the call graph. In fact, a model formulation based on PGMs suited to this particular problem is presented in Appendix A. Unfortunately, it was not possible to include a proof of concept for said method in this thesis but it was nonetheless explored in detail and it presents a very promising approach.

### 5.5 Related work

In their work "Inferring user demographics and social strategies in mobile social networks" [2], Dong et al. approach the problem with a probabilistic graphical model, introducing the usage of factor graphs in order to predict age and gender, simultaneously, from call graphs. In this context, it might be considered the most complete and efficient model developed to date.

The authors constructed an algorithm that supports multiple dependent-variable classification through modeling the interrelations among different dependent variables. They propose a Double Dependent-Variable Factor Graph model to infer users’ age and gender simultaneously, claiming the
Figure 5.13: Performance analysis for gender prediction with three tree-based methods, using 10 folds. A comparison between tests with network related features and without them is presented. There an accuracy improvement for all three methods, when predicting age considering network features. However, it is still a marginal improvement, which may further confirm the unsuitableness of such simple classification method for leveraging the network structure when inferring gender from call graphs.

Figure 5.14: Performance analysis for age prediction with three tree-based methods, using 10 folds. A comparison between tests with network related features and without them is presented. As a matter of fact, accuracy improvements are even less discernable in the case of age prediction using tree-based classifiers. There is a clear lack of capability to capture the network structure when inferring age from call graphs using classification methods.

advantage of exploring not only the relations between demographics and communication behaviors, but also considering the interrelations between the two demographics gender and age. They define the demographic prediction problem: let $G = (V^L, V^U, E, Y^L, Z^L)$ denote the partially labeled network, where $V^L$ and $V^U$ are the set of labeled and unlabeled nodes in the graph, respectively, $E$ is the set of graph edges, and $Y$ and $Z$ correspond to the gender and age variables for labeled nodes. Further considering an attribute matrix $X$, where each row $x_i$ represents the feature vector for user $i$, the objective is “maximizing the conditional probability of users’ gender $Y$ and age $Z$, given their corresponding attributes $X$ and the input network structure $G$, i.e. $P_{θ}(Y, Z|G, X)$”[2]. Factor graphs were used so as to factorize the global probability distribution into a product of local factor functions, allowing the simplification of the objective function. As shown previously in Figure 1.5 the authors define three different types of factors, representing the correlation between the demographic variables $Y_i, Z_i$ and the set of features $X_i$ of user $i$, the dyadic relations between users $i$ and $j$ and the triadic relation among users $i, j$ and $k$. All factors are instantiated by an exponential-linear function. In terms of feature definition, they adopt a set of 24 non-structural attribute features comprising individual features,
friend attributes and circle attributes. Both the dyadic and triadic factors are said to “capture structural features designed to model the demographic distributions over edges and triangles with both labeled and unlabeled users”. When setting up the experiments, they do a separate analysis of the call and sms networks. Both graphs are constructed with fully labeled nodes and one edge is added between nodes \( i \) and \( j \) when there exists at least one reciprocal communication – calls or sms, respectively – between them in the CDRs collected over two months. Non-active users are filtered out and from then on, only the largest connect component of each graph is used when performing inference experiments. In the case of the largest connected component of reciprocal call network, it is composed by around 1.09 million nodes, whilst the largest component of the reciprocal sms network comprises approximately 304 thousand nodes. Experiment predictions are repeated ten times, after which the average performance is reported in terms of weighted Precision, Recall and F1-Measure. Weighted analysis is used since all classes female/male or young/young-adult/middle-age/senior are considered to have the same relevance. Comparisons are carried with different classification algorithms, namely Logistic Regression, Support Vector Machines, Naive Bayes, Random Forest, Bagging, Gaussian Radial Basis Function Neural Network and a simplified Factor Graph Model. 50% of the labeled data in each one of the two networks is used as training set and the remaining 50% is used for testing. The results show an improvement of around 10% compared to the baseline methods for prediction of user’s age and gender. The authors reported accuracies of 80% and 73% for inferring user’s gender in the call network, and for inferring user’s age in the sms network, respectively, which represent the best scenario configuration for the prediction of each demographic property. As reported by the authors, the dataset used in their study was neither noisy nor imbalanced.

The usage of probabilistic graphical models for inference of users’ demographic properties through mobile call graphs seems rather interesting, however the presentation of the work done by Dong et al. in [2] does not clearly expose the formulation of the method nor does it clarify how inference is really done. More importantly, the code provided did not allow for the reproducibility of the results, therefore not allowing for an evaluation of the algorithm’s performance with the dataset available to this thesis. Because of that this model cannot be used as a benchmark. It is indispensable to emphasize that the accuracy of the results heavily depends upon the quality of the dataset.

Based on work done on TRW methods for optimization in MRFs [11], the author of [12] proposes a modification so that the value of the lower bound of the energy is guaranteed not to decrease. In other words, the modification presented as TRW-S, guarantees “to find a “local” maximum of the bound”. Since dual optimization is used, finding the maximum of the lower bound of the energy function – which is a convex function – consists of finding the maximum of its dual function – a concave function [48] – and all “local” maxima of a concave function are global, thus supporting the proposition of better results using TRW-S. In the message passing scheme that serves as base to the algorithm, messages are updated not in parallel as suggested by Wainwright in the original TRW techniques, but in a sequential order. That being said, the TRW-S method was shown to improve over the results of TRW methods, with the work of Kolmogorov. Indeed, in [49] a comparison of the performance of various energy minimization methods on different benchmark problems is presented and TRW-S
“gave consistently good results”.

Another interesting framework for solving MRF-based optimization problems is specified in [10], taking advantage of the powerful technique of dual decomposition. In reality, Komodakis et al. suggest addressing MRF optimization problems, which are NP-hard since they are based on an undirected probabilistic graph $G$, by decomposing them into subproblems, each one defined on a tree $T \subset G$, and then combining the solutions in a principled way. Given a set of labels $\mathcal{L}$ and a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, “the task of MRF optimization then amounts to assigning a label $l_p \in \mathcal{L}$ to each node $p \in \mathcal{V}$ such that [...] the MRF energy function is minimized”. The method is based on the linear relaxation of the energy minimization integer program to which TRW methods are tightly related. As mentioned before, TRW methods do not attempt to minimize the linear relaxation $\text{LP}_{\text{MRF}}$ directly, instead focusing on solving the dual of that relaxation. “Obviously, the cost of any feasible solution to the dual of $\text{LP}_{\text{MRF}}$ yields a lower bound on the optimal MRF energy. Hence, solving the dual corresponds to a maximization of the lower bound”, which is essentially the key idea behind the TRW methods. The framework deployed relies on a widely popular optimization technique named dual decomposition: “given a difficult or large problem, we decompose it into smaller solvable subproblems and then extract a solution by cleverly combining the solutions from these subproblems”. This technique implies defining the subproblems, also known as slave problems, and a coordinator, the master problem. Furthermore, one can either decompose the original problem – primal decomposition – or its Lagrangian dual – dual decomposition. The algorithm will then operate on two levels. On a lower level, slave problems are solved. They consist of MRF optimization problems applied to the tree graphs in $T$, which are easy to solve given their structure. On a higher level, a master problem will make sure to coordinate the slave problems so as to optimize the general MRF problem – it will update the parameters of each slave’s MRF problem at each iteration, based on the results produced by all slave problems in the previous iteration. In order to optimize the master problem, the authors use a projected subgradient method [10], in which the dual variables of the Lagrangian dual function are updated with the objective of approximating the assignments of the individual trees. Since the optimization is solved on the dual space for each tree, the obstacle then lies on finding the optimal primal solutions to the general MAP assignment. The drawback of working on the dual space is that there is no guarantee of being able to recover the optimal values for the primal variables.

If assuming probabilistic graphical models, in particular an MRF formulation of this thesis’ inference problem, such optimization methods can be extremely rich frameworks for predicting the most probable demographic classes of mobile phone users based on mobile call graphs. In addition, it might be necessary to consider tree structured graphs for reliably performing inference tasks. Finally, for complex optimization problems, decomposition into smaller, easier to solve, problems might be the best option.
Conclusions and future work
The aim of this thesis was addressing the problem of inferring user demographics, namely gender and age, from mobile call graphs. Working with a dataset that holds anonymized information about mobile phone users’ patterns of communication, an exploratory data analysis was conducted and tree-based classification methods were tested in order to evaluate the extent to which demographic prediction can take advantage of the mobile phone network structure.

Indeed, several state of the art studies suggest that there is a relation between user’s demographic properties and their calling behaviors within the mobile phone network. By conducting extensive exploratory data analysis on the available dataset, several social strategies were highlighted. To begin with, it was observed that females tend to communicate more often than males, and predominantly with other females. Younger generations are more active in their communications and have a higher predisposition to use sms than older generations, which, on their part, tend to have closer relations among themselves. Evidence was found for gender and age homophily in call graphs, once again respecting the assertions made by state of the art. Moreover, cross-generation and cross-gender frequency of communications denote common social paradigms such as relationships among parents and children or of conjugal character. It was also found that, when forming groups of three connected users, males tend to connect to other males, while females are relatively more likely to take part in different gender triads, being that seniors have a higher presence in social triadic circles, when compared to younger generations. Dong et al. suggest in [2] that young users are more active in broadening their connects, i.e, tend to have higher node degree values, while older users typically have more stable connections, assuming high clustering coefficients. This was not supported in the dataset available to this thesis and it might be so because of cultural differences or it could be a simple artifact of the dataset limitation.

Based on the findings of the EDA, it is proposed that demographic inference from call graphs should leverage the mobile network structure and the behavioral patterns it represents. As such, tree-based classification methods were tested to further evaluate the importance of network metrics. In fact, the results did not demonstrate the assumption: overall the prediction power of tested method over the dataset with and without the network metrics was not high, resulting on average accuracy values of no higher than 62% for gender and 55% for age prediction. Possible factors that impacted these results include the imbalance of the dataset, the noise of gender and age labels or even the lack of more specific information such as communication timings to distinguish between on and off-hours. Unfortunately, the state of the art proposed by Dong et al. was not reproducible, but one should also expect that the accuracy results would depend heavily upon the characteristics of the dataset, which are not the same for their dataset as are for this one.

In conclusion, tree-based classification methods might not be the best way to approach demographic inference from call graphs. Indeed, this work opens the way for the development of a probabilistic graphical model that effectively takes advantage of the graph structure and allows efficient inference of demographic properties of mobile phone users. In particular, Markov Random Fields (MRF) are an extremely interesting way of framing this prediction problem, which can be further solved with a Maximum a Posterior (MAP) approach. Since the MAP assignment problem in MRFs is,
essentially, an energy minimization problem of NP-hardness, a promising model would entail a maximum spanning tree approximation (via energy minimization) of the MRF graph, and consequent exact inference on the tree structured subgraph. When repeated iteratively, such algorithm could improve the tree approximation at each iteration, therefore being able to converge to the minimum value of the energy of the MRF. A detailed theoretical formulation of a model proposed to tackle this problem is presented in Appendix A. Future work would include the implementation of the PGM based algorithm and the acquisition of results.
Bibliography


Theoretical background - Formulation of a PGM
In order to model the demographic inference problem the PGM framework can be used, since it provides a compact and efficient way of representing a multivariate probability distribution and further allows performing inference tasks based on the graphical structure. There are two different types of random variables in the proposed model: $W_i$, representing the class of user $i$, and $X_i$, representing the set of features extracted for user $i$. Considering the 17334 total users in the mobile call network, the probabilistic model will comprise a number of random variables that is double that amount. Furthermore, the aim is to model not only the dependencies among variables $W_i$ and $X_i$, i.e., model the probability distribution of user $i$ being of a class $W_i$ and having a set of characteristics $X_i$. It is also desirable to consider the relations among different users in the mobile call network, modeling the dependencies between $W_i$ and $W_j$.

Figure A.1: Example of a small social network represented by a mobile call graph $G = (V, E)$. This network is composed by $|V| = 9$ users, connected among each other through $|E| = 14$ edges, conditioned upon the CDRs. Users $i$ and $j$ are connected in the graph if they communicated bidirectionally – either calls longer than 5 seconds or sms – at least once in 6 months.

That being said, Figure A.1 will be used henceforth to illustrate the proposed model, referring to a small mobile call network example. The model assumes bidirectional dependencies, meaning that, for example, if user $i$ and user $j$ are connected in the mobile call graph, then the class of user $i$, $W_i$, will influence the class of user $j$, $W_j$, and vice-versa. The independence network is, consequently, represented by an undirected graph $G = (V, E)$ and the joint probabilities of pairs of variables $W_i$, as well as of variables $W_i$ and $X_i$, are modeled as clique potentials. Given that there is a repetition of the

Figure A.2: Template model. The PGM is represented as an undirected graph, comprising two different kinds of random variables associated to each user $i$: $W_i$ represents the class and $X_i$ represents the set of features. For each $X_i$ there is an emission probability distribution $P(X_i)$. In terms of clique potentials, there exist unary potentials $P(W_i, X_i)$ and pairwise potentials $P(W_i, W_j)$. Since the random variable semantics are repeated for all users, a template probabilistic model is used for each user $i$ and the connections between $W_i$ and $W_j$ variable nodes are defined according to the connections between users $i$ and $j$ in the mobile call graph.

random variables’ semantics for each user $i$ in the probabilistic model, a template model was used, as shown in Figure A.2.
Figure A.3: Representation of the model as a Markov Random Field. The independence network consists of an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ whose nodes $Y \in \mathcal{V}$ are the random variables $Y = W_1, ..., W_n \cup X_1, ..., X_n$. It defines the independencies between random variables and encodes the joint probability $P_\Phi(Y)$ as a product of factors $\Phi = \{\phi_1(D_1), ..., \phi_k(D_k)\}$, where $D_i \subseteq Y$.

Essentially, the problem can be formulated as a Markov Random Field (see Figure A.3), where the independence network encodes the pairwise, unary and emission potentials as defined in the template model. This MRF comprises the random variables $Y = W_1, ..., W_N \cup X_1, ..., X_N$, where $N = |\mathcal{V}|$ is the number of users in the network, and defines $P_\Phi$ as a Gibbs distribution parameterized by a set of factors $\Phi = \{\phi_1(D_1), ..., \phi_k(D_k)\}$, where $D_i \subseteq Y$, as follows:

$$P_\Phi(Y) = \frac{1}{Z} \tilde{P}_\Phi(Y),$$

(A.1)

where

$$\tilde{P}_\Phi(Y) = \phi_1(D_1) \times \phi_2(D_2) \times ... \times \phi_m(D_m)$$

(A.2)

is an unnormalized measure and

$$Z = \sum_Y \tilde{P}_\Phi(Y)$$

(A.3)

is a normalizing constant, also referred to as the partition function.

Figure A.4: Formal representation of the model as a Conditional Random Field. The independence network is represented by a partially directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ whose nodes $Y \in \mathcal{V}$ are given by $Y = W_1, ..., W_n \cup X_1, ..., X_n$. The distribution encoded considers a set of factors $\Phi = \{\phi_1(D_1), ..., \phi_k(D_k)\}$, where $D_i \subseteq Y$, $X_i$'s are assumed to always be observed, hence being depicted in grey.

Conditional Random Fields are a variant of MRF for which each random variable can be conditioned upon a set of global observations $O$. The advantage lies on the fact that there is no need to
model the distribution over the said observed variables, which are not the particular focus variables of the problem – the target variables. This framework is better suited for discriminative classifiers, which do not take interest on the distribution of the observations. In this particular model, the main focus is in predicting the probability distributions of $W_i$ variables and so variables $X_i$ can be seen as observations that condition the set of target variables $W_i$. When modeling the problem as a CRF each clique potential $\Phi_k = P(W_i, W_j)$ will be a mapping from all assignments to both the clique $k$ and the observations $O$ that condition the scope of the clique, $W_i$ and $W_j$. A CRF can be represented by a partially directed graph, as depicted in Figure A.4, whose nodes will correspond to the disjoint observed variables $X$ and target variables $W$. "At a high level, this graph is parametrized in the same way as an ordinary Markov network, as a set of factors $\Phi = \{\phi_1(D_1), ..., \phi_k(D_k)\}$" [9], where $D_i \not\subseteq X$.

A conditional probability is encoded such that:

$$P(W|X) = \frac{1}{Z(X)} \tilde{P}(W, X),$$

where

$$\tilde{P}(W, X) = \phi_1(D_1) \times \phi_2(D_2) \times ... \times \phi_m(D_m)$$

is the unnormalized measure and

$$Z(X) = \sum_W \tilde{P}(W, X)$$

is the partition function. More formally, a CRF is represented by an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where two variables are connected by an undirected edge whenever they are included in the scope of the same factor. As it turns out, the graph can be viewed as representing the conditional probability $P(W|X)$ rather than $P(W, X)$, thus avoiding the representation of a probabilistic model over $X$.

Figure A.5 presents an equivalent, simplified independence network of the CRF proposed, where the observed variables $X_i$ are implicitly conditioning the distribution of target variables $W_i$.

Once the model’s representation is established, one can use the PGM to answer queries about the distributions of its random variables. Inference in this model will assume a Maximum a Posteriori approach in which the aim is to find a single coherent joint assignment to the non-observed $W_i$ variables that is the most likely, thus maximizing the conditional probability $P(W|X)$. Formally, the
problem is defined as given a set of observed variables, also known as evidence, $X = x$, the MAP assignment to all non-observed variables corresponds to

$$\text{MAP}(W|x) = \arg \max_w P(W,x),$$

(A.7)

where $w, x \in \mathbb{R}^N$ and $N$ is the number of users in the network, i.e., nodes in the graph. Considering the definition of conditional probability, where $P(W,X) = P(X)P(W|X)$ and bearing in mind that the distribution $P(X) = P(X = x)$ is constant since $X$, are observed variables, the inference objective can be written as

$$\text{MAP}(W|x) = \arg \max_w P(W|x).$$

(A.8)

It is important to notice that a MAP assignment is not the same as maximizing over the marginal distributions. Indeed, “the assignment where each variable individually picks its most likely value can be quite different from the most likely joint assignment to all variables simultaneously” [9]. Actually, there may even be more than one optimal solution to MAP inference queries.

In the context of MAP, given a PGM $P_\Phi$, finding a joint assignment $w$ which maximizes the probability $P_\Phi(W)$ is, in fact, an NP-hard problem. NP stands for Non-deterministic Polynomial-time and is a class used to describe the complexity of some decision problems. What this means is that exact inference – and approximate inference for that matter – in graphical models probably requires exponential time in the worst case (except in the unlikely event that $P = NP$). Nevertheless, one often does not care for the worst case, and rather focuses on practical real-world applications, many of which can be tackled very effectively using exact or approximate inference algorithms for graphical models.

Factor graphs are a finer-grained representation of a Markov network with the advantage of providing a more explicit structure of the probabilistic distribution in some cases. However, factors are essentially complete tables that map all assignments to the variables that compose that factor’s scope, which can still be large and difficult to work with. An alternative parameterization of the factors comes from adopting log-linear models. More concretely, one can rewrite a factor $\phi(D)$ as

$$\phi(D) = \exp(-\epsilon(D)), \quad (A.9)$$

where $\epsilon(D) = -\ln \phi(D)$ is denoted as energy function. This logarithmic representation, which ensures the positivity of the probability distribution, allows for a reformulation of the MAP problem. Given the definition in [A.4], and considering the new parameterization,

$$P(W|x) \propto \exp \left[ - \sum_i \epsilon_i(D_i) \right], \quad (A.10)$$

the inference objective can be rewritten as

$$\max_w P(W|x) = \max_w \exp \left[ - \epsilon(W|x) \right] = \min_w \epsilon(W|x), \quad (A.11)$$

where $\epsilon(W|x)$ is the energy function of $W$ given the evidence $x$. In this particular case of node classification, each node $W_i$ in the graphical model, taken in isolation, has a preferred value for the class to which it belongs. Nonetheless, it is still desirable to impose a soft “smoothness” constraint over the entire graph, so that neighboring nodes should take concordant class values. In MRF and CRF, the
nodes’ preference is encoded in the unary potentials, and the smoothness preferences are encoded in the pairwise potentials. In this sense, the energy function will assume two different terms:

\[ \epsilon(W|\mathbf{x}) = \sum_{(u,t) \in E} \theta_{ut}(w_u, w_t|\mathbf{x}) + \sum_{u \in V} \theta_u(w_u|x_u), \]  

(A.12)

where \( E \) and \( V \) denote the set of edges and nodes in the independence network graph \( G \), respectively. Apart from \( G \), a discrete set of labels, or classes, \( L = \{1, 2, 3, 4, 5, 6, 7, 8\} \) is defined as representing the eight possible demographic classes established (possible combination of gender and age labels) and so \( w \in L \). In the first term, the pairwise potential functions \( \theta_{ut}(w_u, w_t|\mathbf{x}) : L \times L \rightarrow \mathcal{R} \) define the energy of two nodes \( u \) and \( t \) connected by an edge, being of class \( w_u \) and \( w_t \), given the set of evidence \( \mathbf{x} = \{x_u, x_t\} \). In the second term, the unary potential functions \( \theta_u(w_u|x_u) : L \rightarrow \mathcal{R} \) represent the energy of user \( u \), being of class \( w_u \), given the evidence \( x_u \). As such, \( \Theta_u(w_u|x_u) \) is the \([1 \times 8]\) column vector defined as

\[ \Theta_u = \begin{bmatrix} \theta_u(W_u = 1|x_u) \\ \theta_u(W_u = 2|x_u) \\ \theta_u(W_u = 3|x_u) \\ \theta_u(W_u = 4|x_u) \\ \theta_u(W_u = 5|x_u) \\ \theta_u(W_u = 6|x_u) \\ \theta_u(W_u = 7|x_u) \\ \theta_u(W_u = 8|x_u) \end{bmatrix}. \]  

(A.13)

In the same way, \( \Theta_{ut}(W_u, W_t|\mathbf{x}) \) is defined as the \([8 \times 8]\) matrix that holds the energy of all combinations of assignments to the target variables \( W_u \) and \( W_t \), given the set of evidence \( \mathbf{x} = \{x_u, x_t\} \):

\[ \Theta_{ut} = \begin{bmatrix} \theta_{ut}(W_u = 1, W_t = 1|\mathbf{x}) & \theta_{ut}(W_u = 1, W_t = 2|\mathbf{x}) & \cdots \\ \theta_{ut}(W_u = 2, W_t = 1|\mathbf{x}) & \theta_{ut}(W_u = 2, W_t = 2|\mathbf{x}) & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}. \]  

(A.14)

Let \( \Delta \) be the space of all selectors \( \mathbf{e} \) such that

\[ \Delta = \{ \mathbf{e} : e \in \{0, 1\}, \mathbf{e} \mathbf{1} = 1 \}. \]  

(A.15)

It then becomes possible to rewrite the energy function in the matricial form:

\[ \epsilon(W|\mathbf{x}) = \sum_{(u,t) \in E} \mathbf{e}_u^T \Theta_{ut}(w_u, w_t|\mathbf{x}) \mathbf{e}_t + \sum_{u \in V} \theta_u(w_u|x_u)^T \mathbf{e}_u, \]  

(A.16)

and the MAP problem is translated by

\[ \min_{\{\mathbf{e}_u, \mathbf{e}_t\}} \sum_{(u,t) \in E} \mathbf{e}_u^T \Theta_{ut}(w_u, w_t|\mathbf{x}) \mathbf{e}_t + \sum_{u \in V} \theta_u(w_u|x_u)^T \mathbf{e}_u \]  

s.t. \( \mathbf{e}_u, \mathbf{e}_t \in \Delta \).  

(A.17)

From (A.17) it becomes apparent that this is a combinatorial problem. Concretely, \( \mathbf{e}_u \) and \( \mathbf{e}_t \) belong to the selector space, therefore being \( \mathcal{R}^8 \) column vectors with properties according to \( \Delta \). As can be seen, they appear in both terms of the minimization objective, with particular complexity in the first term where both selectors are multiplying by the matrix \( \Theta_{ut} \). That being said, efforts can be done towards the linearization of the problem.

When looking for a linear approximation of the problem, one can assume as goal the isolation of all non-convexity, which essentially lies on the \( \mathbf{e}_i \) variables. Let \( \Theta \in \mathcal{R}^{(N+1)8 \times (N+1)8} \) be the matrix
that entails all $\Theta_{ut} \in \mathcal{R}^{8 \times 8}$ matrices, such that

$$\Theta = \begin{bmatrix}
0 & \Theta_{12} & \Theta_{13} & \Theta_{14} & \ldots \\
\Theta_{21} & 0 & \Theta_{23} & \Theta_{24} & \ldots \\
\Theta_{31} & \Theta_{32} & 0 & \Theta_{34} & \ldots \\
\ldots & \ldots & \ldots & \ldots & \ldots 
\end{bmatrix}. \quad (A.18)$$

Additionally, let $\vartheta \in \mathcal{R}^{8N}$ be the column vector that holds all $\Theta_u$ and $e \in \mathcal{R}^{8N}$ the column vector with all selectors $e_u$.

$$\vartheta = \begin{bmatrix}
\Theta_{u1} \\
\Theta_{u2} \\
\ldots \\
\end{bmatrix} \quad (A.19)
$$

$$e = \begin{bmatrix}
e_{u1} \\
e_{u2} \\
\ldots 
\end{bmatrix} \quad (A.20)$$

Defining

$$\varepsilon = \begin{bmatrix} 1 \\ e \end{bmatrix} [1 \ e^T] = \begin{bmatrix} 1 & \varepsilon_{12} \\ \varepsilon_{21} & \varepsilon_{22} \end{bmatrix}, \quad (A.21)$$

the linearized minimization problem is given by:

$$\min_{\varepsilon} \left\{ \begin{bmatrix} 1 & \vartheta^T \end{bmatrix}^T \varepsilon \right\} \quad (A.22)$$

s.t. $\varepsilon = \begin{bmatrix} 1 \\ e \end{bmatrix} [1 \ e^T]$

$$e_u \in \Delta$$

which can also be written as

$$\min_{\varepsilon} \left\{ \begin{bmatrix} 1 & \vartheta^T \end{bmatrix}^T \varepsilon \right\} \quad (A.23)$$

s.t. $\varepsilon \geq 0$

$$\varepsilon_{11} = 1$$

$$\varepsilon_{12} = \varepsilon_{21}^T$$

$$\varepsilon_{21} = \varepsilon_{22}^T$$

$$\varepsilon_{12} = N$$

$$\text{rank}(\varepsilon) = 1$$

The method adopted is informally referred to as rank relaxation. This “powerful, computationally efficient approximation technique” has been used for a wide range of very difficult optimization problems \[50\] and defends the formulation of such problems in a format similar to the one deployed in \[A.23\]. It facilitates the identification of the fundamental difficulty of the combinatorial problem: the rank constraint $\text{rank}(\varepsilon) = 1$ is a non-convex function. One should note that both the objective function and the remaining constraints are convex. Thus, a linear approximation could be obtained by simply dropping the rank constraint. An extensive, and at first sight redundant, description of the remaining properties that constrain the variables $\varepsilon$ is used so that, when dropping the rank constraint, the approximation is as tight as possible. Convex relaxation methods are also covered in \[51\] and \[48\]. Unfortunately, this
approximation still leaves great challenges to be tackled: it only provides exact inference on tree structured graphs and considering that \( N = 17334 \) the sizes of matrices \( \Theta \) and \( \vartheta \) make it computationally intractable.

All the above considered, it is imperative to adopt a different approximation strategy, looking back into (A.17). Based on the work of Kolmogorov [12] and Komodakis et al. [10], a framework is proposed: judiciously choose a single maximum spanning tree on which to perform inference and work on the primal space to solve the MAP estimation problem based on the CRF formulation illustrated in Figure A.5.

\( T \) is a spanning tree of graph \( G = (V, E) \) if it contains all nodes in \( V \) and a minimum number of edges that allows maintaining graph connectivity. In a spanning tree there is one and only one path from one node to another, since there are no loops, and a graph \( G \) may have more than one spanning tree. Furthermore, one can construct a maximum spanning tree (MST) from a weighted undirected graph. MST algorithms choose the spanning tree’s edges so as to preserve the maximum cost paths, where a cost of a path is given by the sum of the weights of each edge on that path. Edge weight and edge cost can be used in simultaneous to denote the same concept. Figure A.6 illustrates one possible spanning tree of the example independence graph. Because graphs may have a large number of spanning trees, it might not be possible to find them all in polynomial time for real world graphs. Some algorithms are known for being able to list all spanning trees in polynomial time per tree [52]. However, given the complexity of the original graph, working with various spanning trees would mean a computational overhead for this inference problem. Furthermore, in the approach suggested by Komodakis et al., exact inference is performed in several trees by the slave problems and at the end it is necessary to combine all individual assignments to each tree in a single one. As mention in Section ??, this implies working on the dual space, using the Lagrangian dual functions to approximate the assignments of all trees. The problem is finding the primal variables, of course, which is done using methods that cannot guarantee exact results [10].

Therefore, another strategy can be adopted: judiciously find one single MST that best represents the original graph, model the system as a CRF and perform exact inference to obtain the assignments to the networks’ nodes. At each iteration the MST is adjusted according to the estimates obtained.
Figure A.7: Illustration of the method proposed in [4]. When breaking an edge in the graph $G$ to construct a spanning tree $T$, one should still take into account the contribution that the connection had for each of the end point nodes. That being said, node A will receive a contribution $\alpha$ and node B a contribution $\beta$, that respectively translates the connection between node A and node B, from the point of view of each node.

After performing inference. A relevant underlying idea is that breaking an edge of the original graph $G$ will have an associated cost which should be minimized. In addition, if an edge is broken one should still take into consideration the contribution of that connection to the nodes that acted as end points. Figure A.7 illustrates the model.

Figure A.8: Scheme of the proposed algorithm. Weights are initialized for each edge of graph $G$ and from there an MST is constructed. Iteratively, inference is performed on the MST, after which the predictions are used to update the edge weights and create a new MST. A stopping condition is defined for when the predictions have not changed from the previous iteration.

The algorithm proposed is schematically shown in Figure A.8. It is important to emphasize that this model proposes solving the non-convex optimization in (A.17) by dividing it into different steps that are individually solvable in a tractable manner, while working on the primal space. Considering the formulation in (A.17), in order to solve the MAP inference problem, it is necessary to look for an MST which minimizes the energy of the system. More precisely, one should cut off the edges of $G$ that cost the least to be broken, i.e., for which the energy spent on breaking them is minimum, following the approach used in [4]. With that in mind, the first step is the initialization of the graph's edge weights. For that, a Least Squares (LS) method is used, defined as

$$\min_{\alpha, \beta} \| [\alpha 1_T + 1 \beta^T] - \Theta_{ur} \|_F^2$$

s.t. $\alpha \geq 0$

$$\beta \geq 0$$

(A.24)

By minimizing the sum of the squared residuals, the LS method is finding variables $\alpha$ and $\beta$ that best
approximate the energy of the edge \((u, t)\). The optimal solution of the LS minimization on each edge corresponds to the cost of breaking said edge in the graph. Naturally, when constructing the spanning tree, the goal is to maintain the edges that cost the most to be broken. The variables \(\alpha\) and \(\beta\) will represent the contribution that the broken edge had in nodes A and B, respectively.

Once the MST is obtained, it is then possible to define a CRF model from it. One must define both the pairwise potentials and the unary potentials for each edge and node in \(T\), respectively. The pairwise potentials will correspond to

\[
\phi(w_u, w_t) = \exp(-\Theta_{ut}(w_u, w_t|x))
\]

(A.25)

while the unary potentials will be given by

\[
\phi(w_u) = \exp(-\epsilon(w_u|x_u))
\]

(A.26)

where

\[
\epsilon(w_u|x_u) = \Theta_u(w_u|x_u) + \sum_{(u,i) \in \mathcal{G}} \alpha_{u,i} + \sum_{(j,u) \in \mathcal{G}} \beta_{j,u}
\]

(A.27)

is the energy of node \(u\) considering the contributions \(\alpha_{u,i}\) and \(\beta_{j,u}\) in case there exists an edge \((u,i)\) or \((j,u)\), respectively, that were cut off from \(\mathcal{G}\) when constructing \(T\). After defining the CRF, the classes of the nodes in the training set are provided and exact inference is performed, obtaining the predictions for the nodes in the test set. Such predictions are used as class estimates \(e_u\) of users \(u\) that belonged in the test set. The estimates are then used in the following iteration to updated the values of the edge weights and refine the approximation that the MST makes of the network \(\mathcal{G}\). After initializing the edge weights with a LS method in the first iteration, they are updated in the subsequent iterations using a majorization-minimization (MM) method, according to:

\[
\min_{\alpha, \beta} \|\begin{bmatrix} \alpha \mathbf{1}^T + \beta \mathbf{1}^T \end{bmatrix} - \Theta_{ut} \|_F
\]

s.t. \(\alpha \geq 0\)

\[
\beta \geq 0
\]

\[
\begin{bmatrix} \alpha \mathbf{1}^T + \beta \mathbf{1}^T \end{bmatrix} \geq \Theta_{ut}
\]

\[
\mathbf{e}_u^T [\alpha \mathbf{1}^T + \beta \mathbf{1}^T] \mathbf{e}_t = \mathbf{e}_u^T \Theta_{ut} \mathbf{e}_t
\]

(A.28)

MM is an optimization method that will be used to find the minimum of the objective in (A.28), exploiting the convexity of the function. The edge weights will be sequentially updated until the predictions on the MST no longer change comparing to the previous iteration, corresponding to a convergence of the algorithm to the best results it can produce.