Learning About Clients From Call Graph
Beatriz Gamito
beatriz.gamito@tecnico.ulisboa.pt
Intituto Superior Técnico

Abstract—Companies like telecommunication operators have in their hands massive amounts of information about social behaviors of their clients, particularly since people use mobile phones for both their personal and professional communications. Based on a dataset provided by a telecommunications operator with fully anonymized information about their clients’ communications, the objective of this work is to represent the social network of mobile phone users in a graph, extract information on their behaviors and evaluate the value of the mobile call graph and its network metrics in the inference of demographic information such as gender and age. Extensive exploratory data analysis unveiled social strategies that can be associated to the network structure, such as cross-gender and cross-generation patterns, gender and age homophily and particular habits associated to user’s gender and age group. Most of the findings were also supported in state of the art studies using other datasets, and so the correlation between users’ demographic properties and mobile phone behaviors within the network is confirmed. Therefore, a preliminary approach with classical classification methods was tested, in order to validate the added value of network metrics used as features in the demographic prediction, as supported by the exploratory data analysis. Classification tree-based methods were tested with and without network metrics as features, with negligible performance difference. Even though the classification methods did not yield impressive performances, this study serves as a baseline for a future modeling of the demographic inference problem with powerful graphical tools like probabilistic graphical models, which can effectively grasp the structural properties of the mobile phone network.

Index Terms—classifiers, communication behaviors, demographic prediction, mobile call graph, social strategies, probabilistic graphical models, conditional random fields

I. INTRODUCTION
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 there is a new form of capital for businesses: data. The concept of Big Data was then introduced as large amounts of data became available streaming from various sources. Such readily available massive quantities of information can be used by companies in order to get to know their clients better and to be able to provide a more effective and personalized service. However, it is common to have data samples that are not representative enough, in the sense that they may encompass only a small population. One technology that has become ubiquitous, though, is the mobile phone. For 2017, the International Telecommunications Union estimated 103.5 mobile-cellular telephone subscriptions per 100 inhabitants [17]. This means that telecommunications operators have in their hands large amounts of information that can cover different demographic and social classes. Mobile phone communications’ information comes essentially in the format of Call Detail Records (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 mobile call graphs, in which the nodes represent users and the edges represent communications between them. Network science provides a variety of metrics that allow the study of graphs, which can in itself 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.” [13] Nevertheless, demographic information in not always disclosed by clients, even though it is relevant for the companies. 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.

The remaining of this work is organized as follows: in Sec. II some related work is examined, after which the concept of mobile call graph is defined in Sec. III. Extensive exploratory data analysis was performed as the findings are described in Sec. IV, followed by a presentation of the experiments conducted and their results in Sec. V. Finally, conclusion are drawn and future work is discussed in Sec. VI.

II. RELATED WORK
Inference problems are often solved using capabilities provided by fields as Data Science (DS) and Machine Learning (ML), whilst relying on concepts of probability and graph theory.

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 labelled data set. 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 unlabelled dataset. In particular, some ML problems can be formulated as classification problems, where the objective is to classify each instance of the dataset with one of a discrete set of classes. Inferring mobile phone users’ age and gender can be framed
age group. The authors propose two methods for age group

call graphs in order to infer home location, income level and
demographic and behavioural characteristics – within mobile
tendency that humans have to relate to others with similar
mobile call graphs.

Another relevant framework for this problem are Prob-
abilistic Graphical Models (PGM), inasmuch as this graph
based representation allows a compact characterization 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 trans-
parent for human understanding. A PGM can be represented
as a Bayesian Networks, encoding conditional probabilities
using directed graphs or Markov Networks, encoding joint
probabilities using undirected graphs. Such graphs represent
the independencies among random variables that hold in the
probabilistic distribution and also define the factorization of
the distribution into smaller factors, each one over a smaller
set of variables. This sort of graphical structure allows the
probabilistic distribution to be used effectively for inference.
For example, Maximum a Posteriori (MAP) queries look for
a coherent assignment to the non-observed variables of a
probabilistic distribution that is the most likely, given a set
of observed variables.

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. Optimization using such
models often corresponds to the MAP estimation problem
and is usually “directly described as an energy minimization
task” [11]. Wainwright et al. present a technique for optimi-
ization in MRFs – tree-reweighted max-product message
passing (TRW) – which aims to find a maximization of the
lower bound of the energy function, approximated using a
linear programing relaxation. TRW algorithms as proposed
in [18] are not guaranteed to increase the lower bound on
the energy and may even decrease it and so, Kolmogorov
proposes an adaptation in [10] that promises to yield better
predictive results. Furthermore, Komodakis et al. present in
[11] an interesting tree-based framework for solving MRF-
based optimization problems using dual decomposition.

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 con-
ducted on the problem of inferring users’ demographics from
mobile call graphs.

Work done by Wang et al. [19], explored homophily – the
tendency that humans have to relate to others with similar
demographic and behavioural characteristics – within mobile
call graphs in order to infer home location, income level and
age group. The authors propose two methods for age group

inference: Majority Vote (MAJ), which classifies a subscriber
with the most dominant age group of its neighbourhood and
RANK, which 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 neighbour,
achieving accuracies of below 78%. Classification is made
simply by looking to the direct neighbourhood of a node. In [4]
and [15], the authors address the mobile phone user age pre-
diction problem with a Reaction-Diffusion algorithm, further
considering the information arising from second or higher-
level neighbours, achieving performances of 62%. A rather
interesting approach to the problem is proposed by Dong et al.
[6], 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 demographic attributes
is explored.

There is additional literature related to the study of social
behaviours and demographics properties based on mobile call
information, including [7], [16] and [12].

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. On such grounds, the relevance of this
work is confirmed.

III. MOBILE CALL GRAPHS

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

1 shows an example of a graph. The path between two nodes
in a graph is defined as the collection of edges (hops) that
connect one node to the other and the shortest path between
two nodes is the minimum numbers of hops necessary to go
from one to the other. Taking the graph in Fig. 1 as an example,
there are two possible paths going from node A to node E,
A-E and A-B-C-E and the shortest path between A and E has
the size of 1 hop. The node degree is defined as the number
of edges connecting to one node so it is possible to see that
nodes A and B have degree 2, nodes C and E have degree 3 and nodes D and F have degree 1. Furthermore, the diameter of a graph corresponds to the maximum distance between any pair of nodes, considering only the shortest paths. This graph has diameter equal to 3.

In order to understand the communication behaviours of users of mobile phones, one can try to observe them as a social network and convey the relations among them. 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. One then defines an edge condition which is the set of rules that determine whether or not two users (nodes) will be connected by an edge in the mobile call graph. As in most state of the art, it will depend on the direction of the communications (unidirectional or bidirectional, when source and destination communicate both ways), whether calls are distinguished from sms and the limits on number of communications or duration of calls.

Graph theory defines useful properties that can enrich the analysis of mobile call graphs. A graph \( G \) is said to be connected, when all nodes in \( G \) are connected to each other, direct or indirectly. For disconnected graphs, a connected component is defined as the maximal connected subgraph, and each node and edge belong to exactly one connected component. It is unrealistic to expect that real world social networks can be expressed by fully connected graphs and thus, it is common to analyze exclusively the largest connected component of a graph created from a mobile dataset. The diameter of a graph \( G \) is defined as the maximum distance (shortest path) between any pair of nodes in \( G \). A powerful concept is that of degree distribution. 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.” [13]

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. Furthermore, graph centrality metrics measure the importance of nodes within a graph. They include degree centrality as defined previously; closeness centrality, measuring the average length of the shortest path between one node and all the others; betweenness centrality, translating the number of times a node is located along the shortest path between any two other nodes; and eigenvector centrality, which assigns relative scores to nodes in a graph, based on the adjacency matrix, which is a square matrix indicating which vertices are adjacent to each other in the graph. In addition, clustering coefficient is a measure of the degree to which graph nodes tend to cluster together. 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 possible exist between them.

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, as done in [19], [4], [15] and [6]. Indeed, in this work all the previously mentioned network metrics and properties will be considered both for the data analysis, as well as for the feature definition of the inference problem.

One final remark should be made on network sampling. 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 and so network sampling becomes imperative [1]. The purposes to network sampling can be property estimation – use a sampled graph to estimate properties of the larger graph – or property preservation – find a sample that is representative of the large graph – [2]. Sampling techniques can be grouped in node sampling, edge sampling and exploration sampling methods. Node sampling, in particular Random Node (RN) or Random Degree Node (RDN), suit the property preservation goal but fail to preserve network properties, in particular degree distribution. Similarly, edge sampling techniques such as Random Edge (RE) and Random Node Edge (RNE), also have a hard time preserving network properties, especially clustering coefficient and graph connectivity. If the sample size is small, which it normally is, the outcome is generally a sparsely connected graph, with no respect for the original community structure. Conversely, exploration sampling methods, like Random Walk (RW), Snowball Sampling (SS) or Forest Fire (FF), 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.

In the context of this work, the original network's dimension and its consequent complexity constituted a challenge for the testing of inference algorithms. As such, SS was used to produce a sampled network that not only has a connected subgraph, but also preserves the original network's properties. In fact, this sampling technique is commonly used for static network sampling [14].

**IV. EXPLORATORY DATA ANALYSIS**

Leveraging from the DS field, this study proposes an analysis of fully anonymized datasets provided by a telecommunications operator in order to obtain knowledge regarding the communication patterns of the users and how they relate to their demographics. Exploratory Data Analysis (EDA) is a process commonly adopted in many DS projects with the
objective of assessing the quality of the data prior to modeling it.

A. Description of the dataset

Two different, but complementary, fully anonymized datasets were provided for this work 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 removed, identities are unrecoverable. The gender and age labels were obtained by the telecommunications operator with a disclaimer of being noisy. Due to privacy constraints, the sources of such labels were never questioned in the context of this work. The data was collected in 2013 over a period of 6 months. Both datasets were merged, resulting in a mobile call dataset with the crossing between identifiers of phone numbers in the communication records and the associated demographic information in the additional dataset. Some of the intervenients of the communications registered were left unlabelled, since the demographic information associated to such phone numbers’ anonymous identifiers was not available. The gender label is either male or female, while the age label 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 mobile call dataset is composed of approximately 157.8 thousand distinct communication records, and around 50% of the observations have missing information, namely labels for the age and gender variables, on one of the intervenients. 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.

B. 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. Analyzing the distribution of gender and age labels in the dataset, it is clear that there is an overrepresentation of male over female and of older generations over younger generations of mobile phone users, as shown in Fig. 2. This skew in the labels’ distribution is commonly denoted as an imbalance on the dataset and will introduce an important bias factor in the data analysis, influencing the accuracy of the results shown.

From the mean values of the variables number of calls and number of SMS for gender and age groups depicted in Fig. 3, 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. 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.

C. Social strategies

The EDA process done over the mobile call dataset allows for a recognition of certain social patterns. Because there is an overrepresentation of male over female users and of older over younger age groups, most social strategy analyses will be impacted. 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 common 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 – 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 users, yet it is more unclear for female users, given that there are predominantly males in the dataset, as shown in Fig. 4. Still, the percentage of female contacts is higher in the case of female users (Fig. 4a) than of male users (Fig. 4b), which indicates that women tend to connect to other women more than men do. Age homophily is also commonly recognized 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. Instead, one can look at the percentage of each age group, across all graphics, in Fig. 5. In fact, it is noticeable that the percentage of contacts of a given age is bigger for users of that same age. Young contacts are in higher percentage in Fig. 5a in detriment of Fig. 5b, 5c or 5d and the same happens for the other labels.

Another interesting analysis has to do with cross-gender communication behaviours. 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 6a, it is possible to see that females communicate more frequently to other females, than males do with other males. Interestingly, mobile communications between users of different gender are also very frequent. Fig. 6a highlights not only gender homophily in mobile call graphs, but also cross-gender social patterns, possibly associated to conjugal relationships. 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. Fig. 6b shows that users labelled as young communicate with extreme frequency to users labelled as young or young adult. As the generational gap increases among mobile phone users, it is also curious to see patterns of social links presumably between parents and sons. 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.

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 three interconnected nodes. Fig. 7a shows that of the females who are included in labelled 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 (Fig. 7b). From Fig. 8, older generations are the most frequent elements of triads in the mobile call graph, which suggests they have closer relationships with their contacts. Dong et al. suggest in [6] 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.
As discussed, several social strategies and communication patterns can be discerned from the analysis of the dataset, 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 the proposition of this work is to develop a model for inferring mobile phone user demographics from mobile call graphs should leverage from the structure of the network.

V. EXPERIMENTS

The idea of this work 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. Indeed, this problem can be framed as a ML problem of classification: the objective is to classify nodes with a gender and age label and assess to which extent is that prediction improved by network metrics.

A. Description of the problem

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. 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 percentage of labeled 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 and so 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.

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.

B. Classification

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. 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. 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. 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” [8].

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 that takes many training sets from the data, builds separate prediction models for each one and averages the prediction results. 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” [8]. 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.

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 [3]. Furthermore, boosting does not rely on bootstrap sampling, but rather fits each tree to a modified version of the original dataset.

C. Experiment setup

Decision trees, RF and GBT models were implemented for gender and age prediction separately. 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 Synthetic Minority Over-sampling Technique (SMOTE)[5], in which the over-sampling of the minority class is performed by creating “synthetic” examples, as opposed to over-sampling by replacement. Even though SMOTE was used to balance the training set, the test data was left imbalanced to mirror the actual problem. The experiments were 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 previously. Furthermore, 4 fold cross validation was used – where 75% of the data was used for training and the remaining 35% 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. Confusion matrices (Fig. 9) are commonly used in statistical learning as a performance evaluation metric. From the numbers presented in confusion matrices it possible to compute accuracy (1), sensitivity, also known as recall (2) and specificity (3).

\[
\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{TP + TN}{P + N} \quad (1)
\]

\[
\text{SN} = \frac{TP}{TP + FN} = \frac{TP}{P} \quad (2)
\]

\[
\text{SP} = \frac{TN}{TN + FP} = \frac{TN}{N} \quad (3)
\]

D. Results

A first set of experiments was conducted where all 39 features were considered, concretely the communication features, the topological features (including network metrics), friend related features and triad related features. Table I shows the accuracies achieved using each one of the tree-based classification methods for predicting gender and age. In general, using 10 fold cross validation improves the results over 4 fold cross validation, since the algorithms are learning from bigger training sets. The confusion matrices were constructed for all prediction scenarios as defined in Fig. 9. In Fig. 10, it is possible to see that, using the Decision tree method for gender prediction with 10 fold cross validation, which corresponds to the most successful experiment, the true distribution of gender labels on the test set is well mirrored: it predicted correctly much more male instances than females ones. For this method, the variable importance plot is presented in Fig. 11. Indeed, network metrics, namely the entropy of sms contacts, the total number of triads and clustering coefficient are among the five most important predictors. However, through all experiments, communication features outnumbered network metrics in the variable importance plots. From the experiment results, such simply constructed classification methods seem to have a hard

<table>
<thead>
<tr>
<th></th>
<th>Gender prediction</th>
<th>Age prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 fold</td>
<td>0.60639</td>
<td>0.62040</td>
</tr>
<tr>
<td>10 fold</td>
<td>0.40781</td>
<td>0.41579</td>
</tr>
<tr>
<td></td>
<td>0.48979</td>
<td>0.50204</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.59848</td>
<td>0.60082</td>
</tr>
<tr>
<td></td>
<td>0.52805</td>
<td>0.53226</td>
</tr>
<tr>
<td></td>
<td>0.55198</td>
<td>0.55094</td>
</tr>
</tbody>
</table>

Fig. 10. Confusion matrix for the best prediction performance: gender prediction with Decision trees, using 10 fold cross validation.

Fig. 11. Variable importance plot for the best prediction performance: gender prediction with Decision trees, using 10 fold cross validation.
time capturing all the information steaming from the network structure.

Another set of experiments was conducted without considering network metrics as classification features, in order to assess how informative those really are for the tree-based methods. A performance comparison is presented for the classification methods that achieve the best accuracies in gender and age prediction, when considering and when disregarding the network related features. For gender prediction (see Fig. 12a) there is a marginal improvement on accuracy when considering network metrics, but in the case of age prediction (see Fig. 12b) 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 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, for example, if considering that professional communications happen predominantly during day time, while night time communications tend to be of personal character.

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 was developed. Unfortunately, it was not possible to include a proof of concept for said method in this work but it was nonetheless explored in detail and it presents a very promising approach.

VI. 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.

Based on the findings of the EDA, it is proposed that demographic inference from call graphs should leverage from 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 without additional features. 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 Posteriori (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 was already done and future work would include the implementation of the PGM based algorithm and the acquisition of results.

REFERENCES