Prediction of Telephone User Attributes Based on Network Neighborhood Information

Carlos Herrera-Yagüe and Pedro J. Zufiria

Depto. Matemática Aplicada a las Tecnologías de la Información,
ETSI Telecomunicación, Universidad Politécnica de Madrid, Spain
carlos@hyague.es, pedro.zufiria@upm.es

Abstract This paper addresses the problem of predicting several attributes corresponding to telephone users, based on information gathered from the network which defines their communication patterns. Two approaches are compared which are grounded on machine learning techniques: the initial approach makes use of link information between two users, looking for the correlation between user attributes and communication patterns. The second approach exploits the network structure underlying the communication behavior of the user under study. Simulations show that the learning machines are able to extract network information to improve the attribute prediction capabilities.

1 Introduction

Communication records between human beings are probably one of the largest source of data generation nowadays. According to recent studies, every minute 370,000 Skype conversations, 198 million e-mails and over half million facebook comments are made worldwide[1]. In addition, only in the US, 2.4 billion calls and 6.1 billion SMSs take place daily[2].

This huge amount of information involves new challenges in the fields of data mining and machine learning. Classical approaches to the study of communication networks consist on aggregating information by user, such as the number of phone calls a user makes, or the time a user calls more often. However, recent works suggest that the aggregation of data may discard a crucial source of information and a new insight for a better characterization of systems: the network structure behind the connections. Many real world systems whose growth was not driven by any pre-existing blueprint have been proved to show non-trivial features when analyzed as a network. Examples include Internet routers [3], web hyperlinks, power grids, neurons connections on simple organisms [15], gene regulatory networks [2] and many others. Social interactions, which can be studied through the proxy of digital communication records, are no an exception for this. Social networks have been found to have a very high number of triangles (referred in literature as the clustering phenomenon), a short diameter and a very

[1] Source: http://www.go-globe.com
[2] Source: http://www.deadzones.com

P. Perner (Ed.): MLDM 2012, LNAI 7376, pp. 645–659, 2012.

heterogeneous degree distribution: there are key nodes, called *hubs,* which have many edges adjacent to them [1]. Besides, it has been found that the structure has a very important influence on the events taking place in the network. For example, Granovetter's hypothesis made back in the 70s [5] turned out to be true according to a recent work by Onella et al. [10] using massive mobile phone data: most of social interaction events happen in small and dense zones of the graph.

The approach used in this work consists of aggregating information by links, and extracting features from both the dynamics of the events in each link and the network structure around the node (user) of interest. The precise problem which will be discussed in this article can be described easily looking at figure 1: given a number of nodes whose attributes are known (well-known nodes, white in the figure) and some links whose events are known, the problem consists of predicting the attributes of an opaque node (black in the figure).

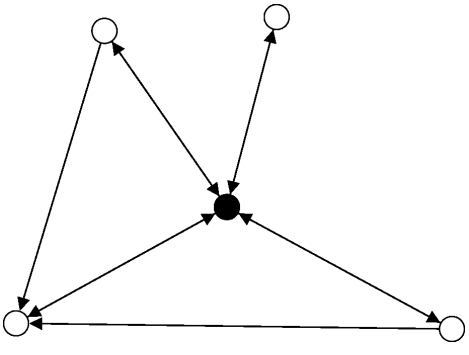

Fig. 1. Problem description: given a number of nodes whose attributes are known (well-known nodes, white) and some links whose events are also known, the attributes of an opaque node (black) must be estimated

2 Some Interesting Scenarios

The problem addressed here does model, in fact, a growing number of real world situations. In many communication services, gathering permission from a user allows the building of his full ego-centered network. This is so because a given link data set can be accessed provided either of the two users involved in the link has granted permission. Some of these situations are described here:

- Facebook applications: if an application is accepted by one user, this user becomes a well-known node, and all its neighbors become potential opaque nodes.
- Twitter private profiles: these profiles would be the opaque nodes, whose ego-networks can be built using their public profile neighbors.
- Mobile phone network: in mobile communications, carriers have information about their customers, and they also have communication records of any interaction between their customers and the rest of the users in the mobile

phone network. This way it is possible to build, using records from only one carrier, ego-networks where non-subscribers are the opaque nodes, and subscribers are the well-known nodes. This scenario will be the case test for this article.

3 Data Description and Preparation

Anonymized data for this article was provided by Orange, France Telecom Spain, consisting of two data-sets:

- Call Detail Records (CDR), contain information about customers interactions via phone calls and SMSs. For each interaction, two anonymized user identifiers and a time-stamp are provided. For phone calls, duration is also available. These data include interaction between subscribers during a continuous 14 week period. According to regulator data, for the observed period, the carrier owned 20% of mobile lines in the country, where mobile phone had already reached market saturation (1.16 mobile lines per person). The data contain records for 2.2 billion interactions among 11 million users.
- User Data (UD): provide age and gender for 8 million users, identified by consistent anonymized hashes.

In order to aggregate this information in a loseless way, records were grouped by relationship (link). For phone calls, along with first and last interaction timestamps, 3 vectors per relation were built:

- Duration vector: contains durations (in seconds) of all phone calls between the two users.
- Inter-event vector: contains inter-event time (in minutes). If the previous vector has length N, inter-event vector has length $N - 1$.
- Direction: binary vector providing caller and receiver roles for the interaction. If the relation is defined as "A-B", this vector has ones when the caller is A and zeroes otherwise.

Once aggregated, 168 million different relationships were found. The next step consisted in filtering meaningful relationships. Many CDR records belong to corporate phone lines for which it does not make sense to talk about user age or gender, since usually there are several persons behind those specific numbers. In this research, the criterion employed to filter out this kind of interactions was the criterion proposed in a seminal paper by Onella et al. [10]: only relationships with at least one call in each direction of the communication were considered. This way, 40 % of originally obtained relations were eliminated from the study.

After having chosen these mutual phone calls relationships, SMSs, inter-event and direction vectors for those relationships were built. The reason for using calls to define meaningful relationships instead of using SMS records, is that the mutual strategy does not work so well for SMS; this is so because of the increasing number of value added services which involve texting in both directions between the user and the corresponding automatic services.

4 Exploratory Analysis and Learning Approach

The resulting network metrics do agree with the metrics proposed in previous works on mobile phone networks [10,6]: high clustering coefficient, short diameter and a long-tail degree distribution. In our data, the degree distribution was found to fit a power-law with exponent $\gamma = -6.27$ which means that there is no scaling behaviour, but the decay does not fit an exponential model either. These features on the degree distribution were also found in previous works, where it was argued that the absence of scaling may be produced by the removal of non-mutual links.

From an exploratory perspective, an interesting behaviour was found in the inter-event time distribution. Basic network engineering does usually assume that the time between two calls (within a big enough population) can be described by a Poisson distribution, and that is the original assumption for many techniques used to properly dimension communication networks [12]. However, recent studies [8,11] have proved that this poissonian behaviour is not present in the individual level: certain events (for example an incoming call) make the user to immediately initiate a communication burst. In our research, once the inter-event vectors for links were compiled, the distribution at this link level was studied.

Inter-call vectors for 10000 randomly chosen links were concatenated, producing 105174 inter-call time samples. Figure 2 shows an histogram of those samples in the time interval from 6 hours to 4 days. One can easily identify that the distribution is peaked every 24 hours. This means that if two users A and B talk to each other by phone today at 10am, it is much more likely for their next conversation to happen the day after tomorrow at 10am than tomorrow at 6pm. This fact contradicts the Poisson distribution monotonic decay, proving that, at link level, there is no poissonian behaviour either.

After the exploratory analysis, a methodology to address the prediction problem was designed, consisting of two separate steps:

- Isolated link prediction: given a relationship A-B, extract features from link available data in order to predict age and gender for B.

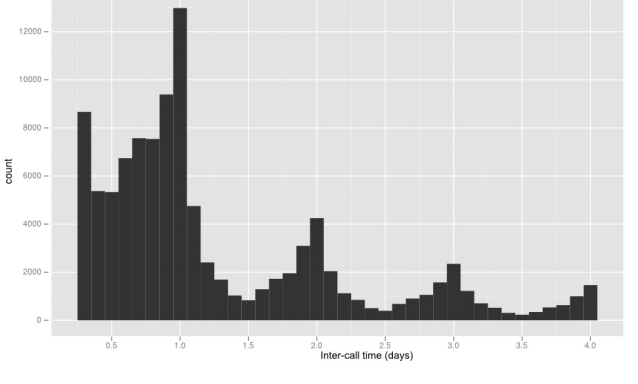

Fig. 2. Inter-call time distribution for a 10000 links sample

– Ego-network: use the results from the previous step to provide a prediction from each link leading to user B. These results, together with ego-network features, are then employed to predict attributes of B.

5 Isolated Link Approach

As previously stated, this section is aimed to perform a most accurate prediction of user attributes (gender, age) using only link related information. Link data gather three main sources of information: SMS records, Call records, and gender/age from the other user in the link.

As it will be shown later, a large part of prediction capability will come from demographic information pertaining to the other user. A reason for this to happen is the well-known phenomenon of homophily in social networks [7], formally defined as the tendency of connected nodes to be correlated. In our stage, this means that if user A and B are connected by a link, it is likely that they both have similar age and even same gender.

Apart from taking advantage of homophily, the manner people communicate also depends on their age and gender, as it was proved by Stoica et al. in [14]. That research found out that it is possible to cluster users into a number of groups according to their communication patterns (unsupervised learning). Then it was shown that some user attributes, such as age, were correlated with the group membership of the user. Although the phenomenon leading to this research is exactly the same, our approach will be grounded on supervised learning (precisely, a classification problem). On the other hand, a larger number of features will be used for machine learning, specially those related to communication dynamics whose relevance has been recently pointed out.

To run this experiment, 9860 links were randomly chosen among those whose both users data (gender and age) were available.

5.1 SMS and Call Metrics

As it has been already mentioned, SMS are one of our three sources of link related data. For each relationship 3 vectors are available which contain all the information associated with direction and communication times. Seminal research on mobile social networks [10] did commonly characterize the link using only quantitative information, such as the number of interactions or total conversation time. Later is was pointed out [8] that, due to the bursty pattern of individual communication, quantitative information may not be enough to describe the nature of the link: 50 messages a week in a relationship may not be more relevant than 10 messages during a month. Using these criteria, new selected metrics from SMS have been calculated:

– Number of SMS during the observation period.
– Mean time between messages and conversation length (from first message to last).

- Variation coefficient (average/standard deviation) for inter-event time.
- Reciprocity: in a link A-B, where B is the opaque node, fraction of messages sent by A. This is the only asymmetric feature for SMS data.
- Fraction of calls during weekend, during work hours, and peak hour of the conversation (0-23).

Concerning call records, the extracted metrics follow similar criteria:

- Number of calls, average call duration, and average inter event time.
- Inter-event variation coefficient and call reciprocity.
- Fraction of calls during weekend, during work hours, and peak hour of the conversation (0-23).

5.2 Gender Prediction

Figures 3 and 4 show the kernel density functions for the metrics described above, aggregated by gender. In general, few gender differences can be found by looking at those graphs; nevertheless, there are a couple of interesting facts to be pointed out. The average call length seems to be higher if user B is female. The median[3] of call duration if B is a female is 89.67 seconds and 75.06 seconds if B is a male. On the other hand, if there is a user sending 20% or less of the messages in the relationship, it is more likely this person to be a man.

Machine Learning Procedure. The problem is defined as a binary classification one. In order to gather an accurate idea of the data quality regarding gender prediction, several learning schemes are tested: Linear Discriminant Analysis (LDA), Decision Trees (Tree), Multilayer Perceptron (Nnet), Bagging and Support Vector Machines (SVM)[4]. In order to improve performance quality (mostly for LDA, which cannot perform non-linear transformations), long tail distributed metrics are logarithmized and, after that, all metrics are standardized (zero mean and unit variance).

Once the data are ready, predictions are obtained from a 10-fold cross validation scheme, ensuring every sample belongs to both training and test sets. Data are provided to the learning machinery in 3 different steps: first only SMS data are provided, then call data are also included, and finally user data are included as well. This whole procedure is summarized in table 1.

The results show an increasing performance, specially when user-data are included. This means that homophily definitely plays a role in this problem. However SMS and call metrics also help to reach a more accurate prediction of user's gender. On the other hand, the capability of splitting non linearly separable sets does not seem to help at all, since LDA performance is almost the same as Nnet or SVM.

[3] Due to the long tail distribution on call duration, it is more robust to use the median to characterize group differences, since the mean is severely biased by outliers.

[4] Implementations used in this article were the following R-packages: MASS (LDA), rpart (Tree), randomForest (Forest), nnet (Multilayer Perceptron), ipred (bagging) and e1071 (SVM).

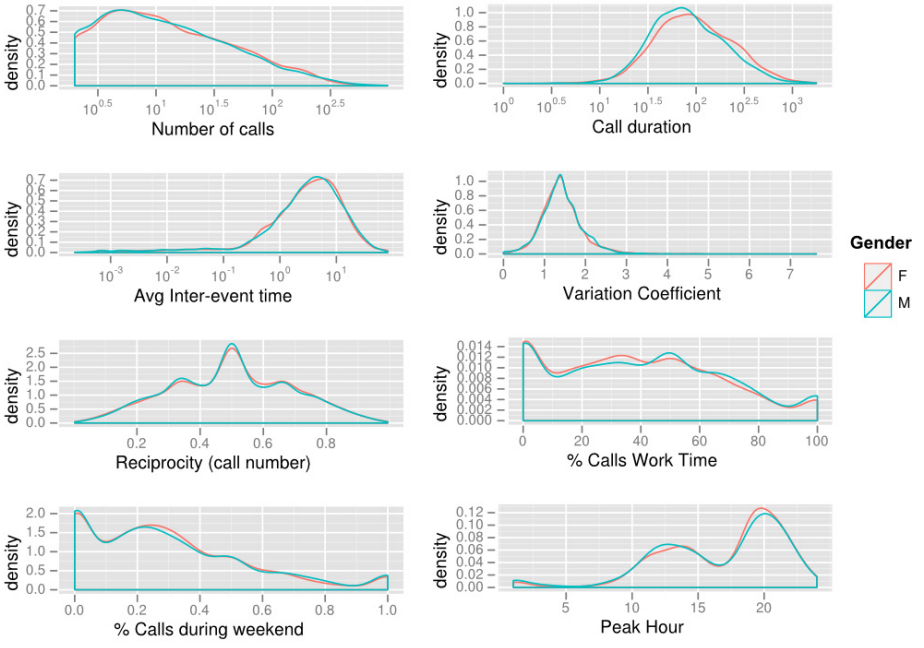

Fig. 3. Density functions for call metrics by gender

Table 1. Gender prediction accuracy using isolated link information

	Random	LDA	Tree	Nnet	Bagging	SVM
SMS	0.5	0.5272	0.5206	0.5334	0.5195	0.5321
SMS +Calls	0.5	0.5403	0.54429	0.5502	0.5343	0.5496
SMS + Calls + User-data	0.5	0.5914	0.60439	0.5945	0.5787	0.6050

5.3 Age Prediction

For prediction purposes, the ages of the users were binned into 6 different age segments. These segments were chosen according to the age distribution, so that every segment has the same number of users in a random sample. This way the age regression problem is transformed in a multi-class classification problem with balanced classes.

After redefining the problem as a classification one, the same methodology discussed in section 5.2 can be applied. Figures 5 and 6 show the density functions for different age groups. An exploratory study shows that people over 30 years old usually call more often during work time. Concerning the amount of SMS in the relationship, it is interesting to note that density functions are sorted, meaning that the elder a person is the fewer texts he/she sends. This behavior was also observed in [14]. However, the statement just made (younger implies more SMS) has an exception in our data which does not show up in any previous

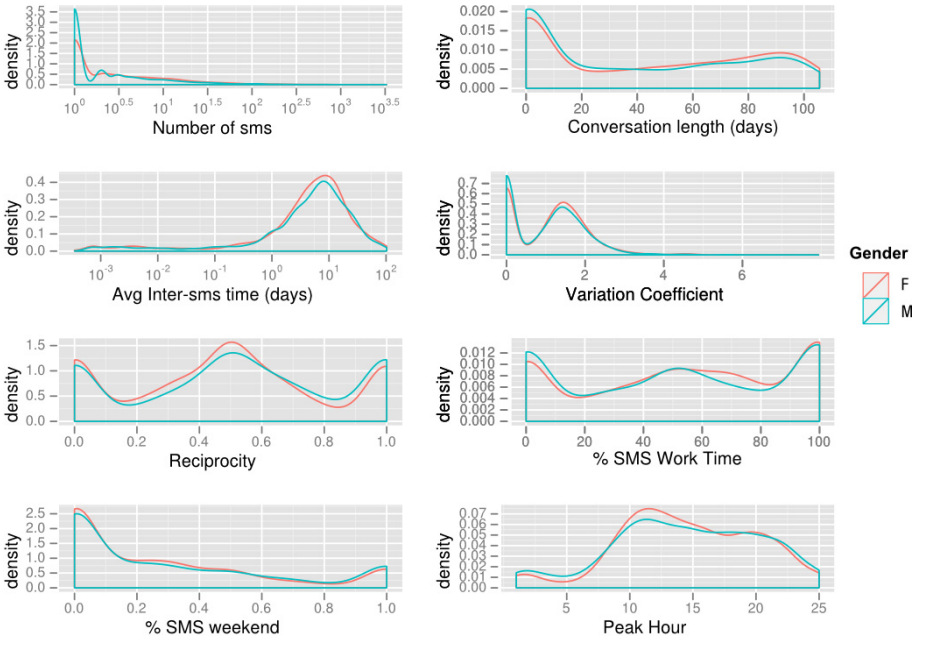

Fig. 4. Density functions for SMS metrics by gender

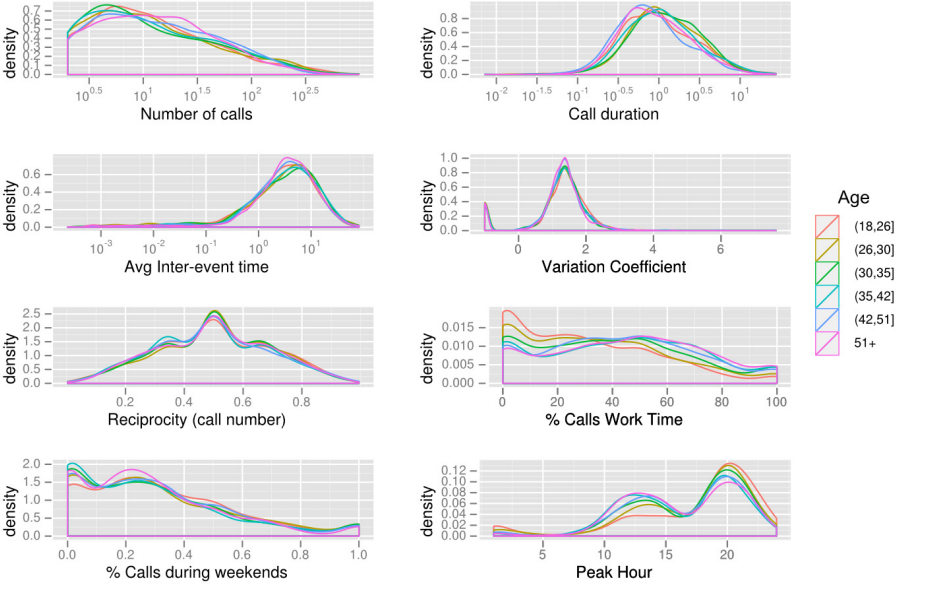

Fig. 5. Density functions for call metrics by age group

Table 2. Age prediction accuracy in using isolated link information

	Random	LDA	Forest[5]	Nnet	Bagging	SVM
Calls	0.1667	0.1997	0.1849	0.2194	0.2020	0.2156
SMS+Calls	0.1667	0.2340	0.2273	0.2410	0.2192	0.2395
SMS + Calls + User-data	0.1667	0.3907	0.4095	0.4027	0.3904	0.4021

study: youngest people (18-26) text a little bit less than people in the next segment. We propose an explanation for that fact: according to recent reports [4], the youngest people (18-25) acquisition of mobile Internet flat rates is higher than in any other age segment. In the same report it is stated that the increase of mobile data plans is severely correlated with the decrease of SMS usage. Hence, we conclude that the observed "lack of messages" among youngest users is probably masked by the replacement of IP-based messaging services whose traces are not recorded by the carrier. The impact of these new technologies on mobile network data should be taken into account in future studies.

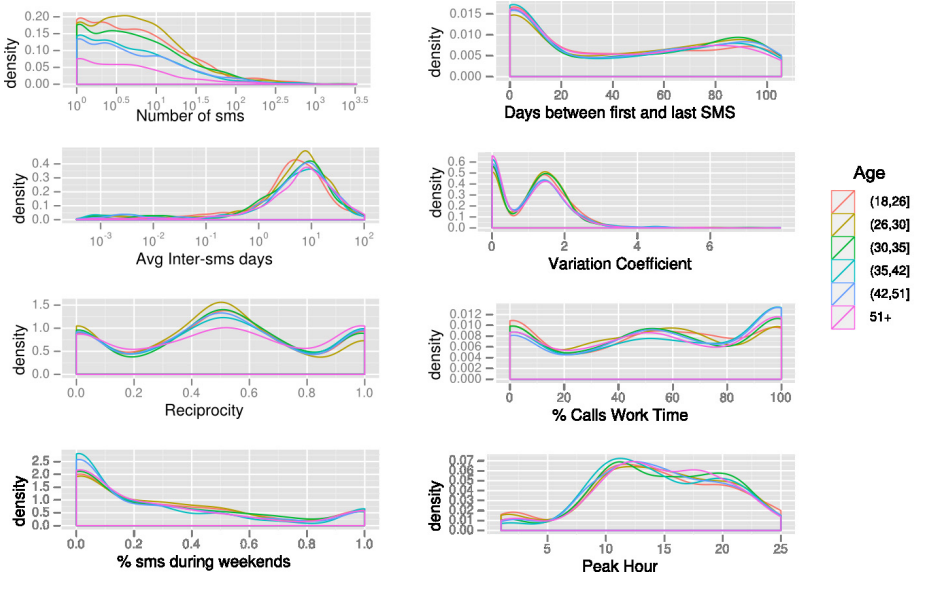

Fig. 6. Density functions for SMS metrics by age group

Table 2 shows the accuracy results for this classification problem. The results show that there is a prediction capability on communication metrics specially if SMS data are included. However, this prediction capability is outperformed if user data (precisely, user age) are included. Age homophily in mobile phone communications is so intense that all five classification techniques mimic the identity function on user age like the best possible classification scheme. The reason for this fact can be observed in figure 7, which represents a scatter density

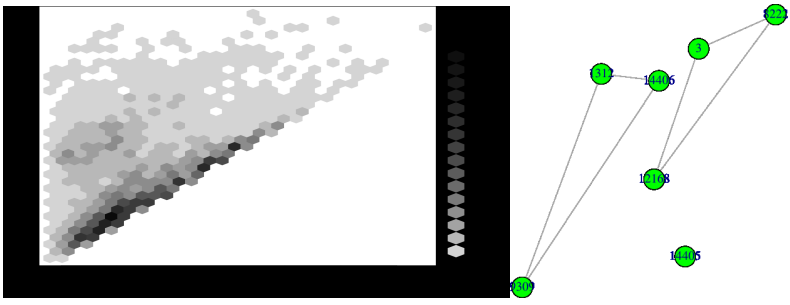

Fig. 7. Left: Age homophily in communication links. Right: ego network: the removal of the opaque node may lead to a not connected graph.

plot for ages in the same link. It is straightforward to check that the probability for a user A to be in the same ages than the user B is extremely high (dark colors in the diagonal of the plot).

6 Ego-Network Approach

The importance of network analysis has been raised over the last decade, where seminal papers by Watts-Strogatz [15], Barabási-Albert [1] and Newman [9] pointed out that many real world systems can be modeled as a network, and the study of the resulting graph usually provides non-obvious and relevant system features. For this reason, this research on multi-link prediction is not only about combining results from single link results, but also including information of the network structure around the opaque node.

6.1 Network Metrics

Due to the size and inherent complexity of their analysis, real-world big networks (mobile phones, social connections online...) are usually analyzed using global information. A very common approach to network analysis is the extraction of a certain N-grade neighborhood around a set of nodes of interest. Among these local neighborhoods, the one which has been more commonly studied is the ego-centered network. This graph includes, for a certain node, all its neighbors and the connections among them. It does not include the center node itself, neither the connections from it to the neighbors, so it is possible a ego-network not being a connected graph, as we can see in figure 7.

Once the subject under study is defined as the ego, a number of features are defined. Apart from quantitative information, such as the number of nodes and edges, some small structures are analyzed. It has been proved that some subgraphs show up much more often than in a purely random network. These subgraphs are called motifs, and their importance in biological networks has been already stressed: the appearance of some kind of motifs is related to some specific

Table 3. Gender prediction accuracy using ego-network data

	Random	LDA	Forest[6]	Nnet	Bagging	SVM
Network	0.5	0.5324	0.5295	0.5362	0.5198	0.5373
Net + Gender Score	0.5	0.6304	0.6323	0.6512	0.6420	0.6532

function within the cell. In social communications networks, the appearance of motifs has also been studied [13,14,16].

According to these guidelines the following metrics were used as network features:

- Node, edge and isolated node count.
- V-motifs (unclosed triangles), closed triangles and 4-star motifs (one node connected to 3).

For this experiment, data from 5670 subscribers were randomly chosen, and their neighborhood information was gathered. This way, a total number of 22098 users and 50377 relationships were analyzed for prediction purposes.

6.2 Gender Prediction

Figure 8 shows that the main difference between gender, regarding network features, is that women seem more likely to have triangles (less stars) in the ego-network. In order to perform final learning for gender classification, results from link level prediction are grouped in a gender score, whose value is the rate of female predictions for the node under study. For example, if there were 3 links, 2 predictions were male and 1 female, the gender score is 0.33. Therefore, a total of seven metrics (six network metrics plus gender score) were analyzed.

Accuracy results are shown in table 3, which shows that the multi-link level increases the performance by about 5% compared to predictions using only one link. On the other hand, figure 9 shows the Receiver Operating Characteristic (ROC) which shows how the accuracy of the best techniques (SVM and Multi-Layer Perceptron) is robust to small variations of the selected threshold.

6.3 Age Prediction

Regarding age, there seems to be a larger diversity in ego-network structure. Figure 10 shows the correlation between age and the appearance of certain motifs, specially stars and triangles. For age prediction, isolated link results were included in the experiment by incorporating 6 variables which contain the number of link level predictions for each label. Prediction accuracy results using these 12 variables (6 age scores and 6 network metrics) are shown in table 4.

Classification results show that the use of network metrics improves link-level predictions by around 10 %, reaching a final performance three times higher than with a random predictor. In addition, prediction errors usually lead to either the

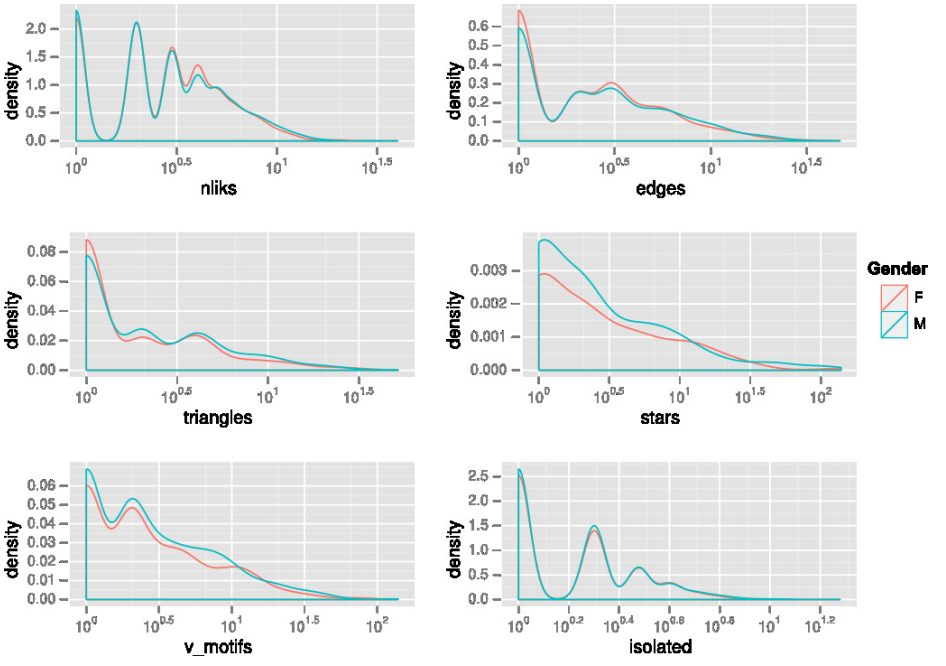

Fig. 8. Density functions for network metrics by gender

Table 4. Age prediction accuracy using ego-network data

	Random	LDA	Forest	Nnet	Bagging	SVM
Network	0.1666	0.2108	0.2022	0.2194	0.2030	0.2092
Net + Age Score	0.1666	0.5050	0.4904	0.5082	0.4931	0.5102

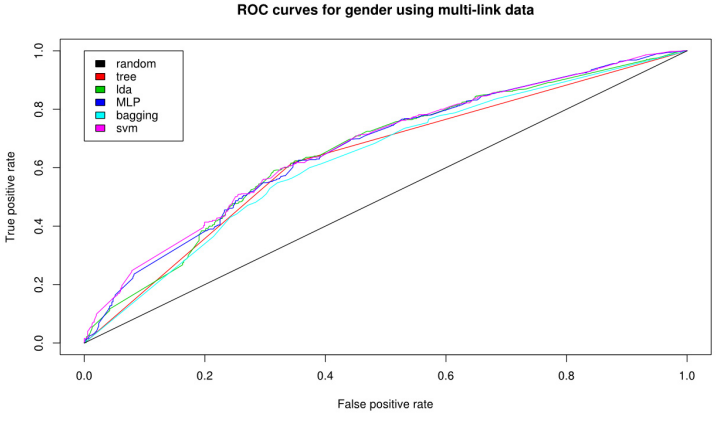

Fig. 9. ROC curve for different machine learning techniques

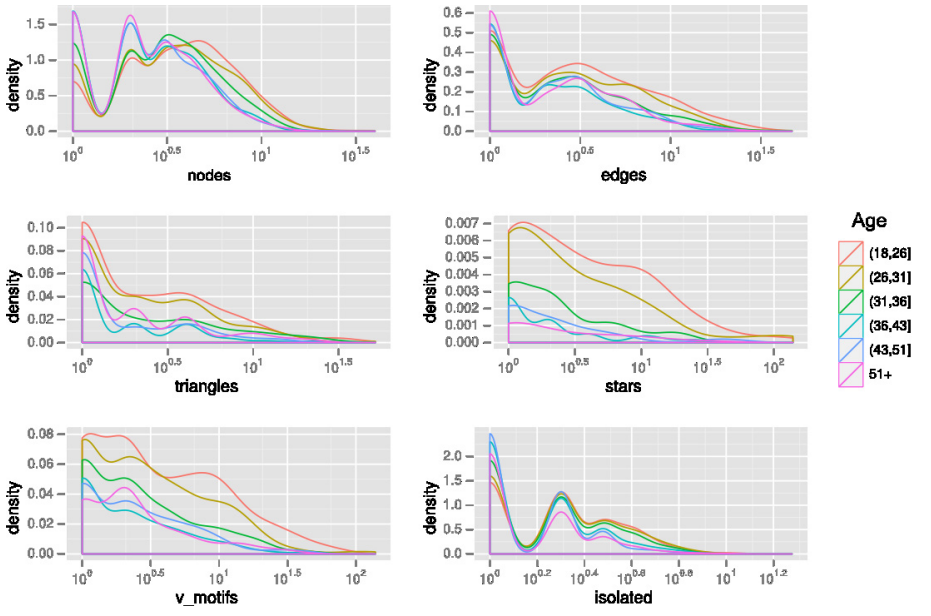

Fig. 10. Density functions for network metrics by age

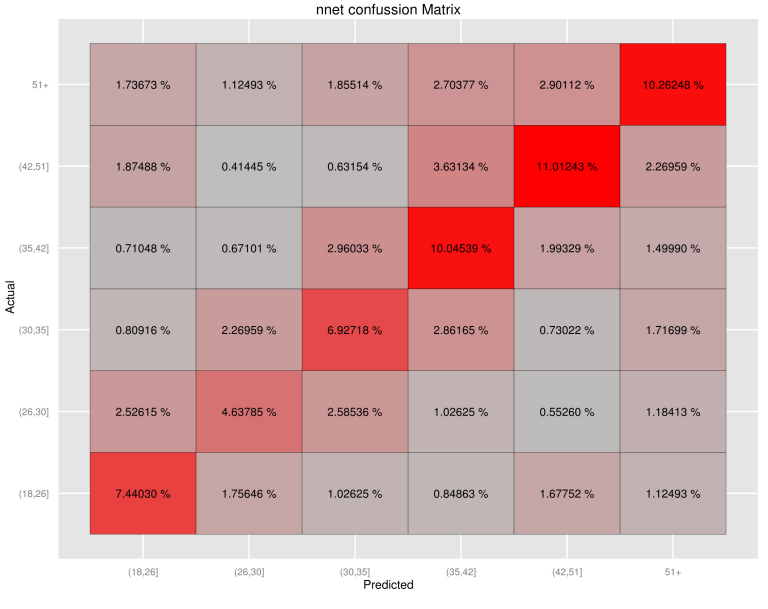

Fig. 11. Confusion matrix for neural network multi-link classifier

following or the previous age element, as it can be seen in the confusion matrix of figure 11. Note that when turning the regression problem into a classification one, the topology on the age variable was neglected; this could have propitiate that errors would lead to age segments non-contiguous with the correct one. Fortunately, the favorable results discarded this main concern.

7 Conclusion and Future Work

The paper has shown that machine learning tools can be a very useful for predicting several attributes corresponding to telephone users, using network data. Two approaches have been evaluated which make use of link information and network structure to improve the attribute prediction capabilities. Future research work on network science may furtherly benefit from machine learning paradigms to generalize the proposed feature extraction schemes when dealing with other different types of networks.

Acknowledgements. The authors want to acknowledge the financial support of Orange (Spain and France), in the framework of Cátedra Orange at the ETSI Telecomunicación in the Universidad Politécnica de Madrid (UPM). The work has been also partially supported by projects MTM2010-15102 of Ministerio de Ciencia e Innovación, and Q10 0930-144 of the UPM, Spain.

References

1. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Reviews of Modern Physics 74(1), 47 (2002)
2. Davidson, E., Levin, M.: Gene regulatory networks. Proceedings of the National Academy of Sciences of the United States of America 102(14), 4935 (2005)
3. Faloutsos, M., Faloutsos, P., Faloutsos, C.: On power-law relationships of the internet topology. ACM SIGCOMM Computer Communication Review 29, 251–262 (1999)
4. Gimeno, M., Villamia, B., Suarez, V.: eEspaña, Informe anual sobre el desarrollo de la sociedad de la información en España. Fundación Orange (2011)
5. Granovetter, M.S.: The strength of weak ties. American Journal of Sociology, 1360–1380 (1973)
6. Hidalgo, C.: The dynamics of a mobile phone network. Physica A: Statistical Mechanics and its Applications 387(12), 3017–3024 (2008)
7. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: Homophily in social networks. Annual Review of Sociology, 415–444 (2001)
8. Miritello, G., Moro, E., Lara, R.: Dynamical strength of social ties in information spreading. Physical Review E 83(4), 3–6 (2011)
9. Newman, M.E.J.: The structure and function of complex networks. SIAM Review 45(2), 58 (2003)
10. Onnela, J.-P., Saramaki, J., Hyvonen, J., Szabo, G., Menezes, A.D., Kaski, K., Barabasi, A.-L., Kertesz, J.: Analysis of a large-scale weighted network of one-to-one human communication. New Journal of Physics 9(6), 25 (2007)

11. Rybski, D., Buldyrev, S.V., Havlin, S., Liljeros, F., Makse, H.: Scaling laws of human interaction activity. Proceedings of the National Academy of Sciences of the United States of America 106(31), 12640–12645 (2009)
12. Schwartz, M.: Telecommunication networks: protocols, modeling and analysis. Addison-Wesley Longman Publishing Co., Inc. (1986)
13. Stoica, A., Couronne, T., Beuscart, J.S.: To be a star is not only metaphoric: from popularity to social linkage. In: Proc. ICWSM 2010 4th. Intl. Conf. Weblogs & Social Media (2010)
14. Stoica, A., Smoreda, Z., Prieurb, C., Guillaumec, J.L.: Age, gender and communication networks. In: Proceedings of the Workshop on the Analysis of Mobile Phone Networks, Satellite Workshop to NetSci. 2010 (2010)
15. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature 393(6684), 440–442 (1998)
16. Zhao, Q., Tian, Y., He, Q., Oliver, N., Jin, R., Lee, W.C.: Communication motifs: a tool to characterize social communications. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp. 1645–1648. ACM (2010)