Inference of Socioeconomic Status in a Communication Graph

Martin Fixman[1,2], Ariel Berenstein[2], Jorge Brea[2], Martin Minnoni[2], and Carlos Sarraute[2]

[1] FCEyN, Universidad de Buenos Aires, Argentina
[2] Grandata Labs, Argentina

Abstract. In this work, we examine the socio-economic correlations present among users in a mobile phone network in Mexico. First, we find that the distribution of income for a subset of users –for which we have income information given by a large bank in Mexico– follows closely, but not exactly, the income distribution for the whole population of Mexico. We also show the existence of a strong socio-economic homophily in the mobile phone network, where users linked in the network are more likely to have similar income. The main contribution of this work is that we leverage this homophily in order to propose a methodology, based on Bayesian statistics, to infer the socio-economic status for a large subset of users in the network (for which we have no banking information). With our proposed algorithm, we achieve an accuracy of 0.71 in a two-class classification problem (low and high income) which significantly outperforms a simpler method based on a frequentist approach. Finally, we extend the two-class classification problem to multiple classes by using the Dirichlet distribution.

1 Introduction

In recent years, we have witnessed an exponential growth in the capacity to gather, store and manipulate massive amounts of data across a broad spectrum of disciplines: in astrophysics our capacity to gather and analyse massive datasets from astronomical observations has significantly transformed our capacity to model the dynamics of our cosmos; in sociology our capacity to track and study traits from individuals within a population of millions is allowing us to create social models at multiple scales, tracking individual and collective behavior both in space and time, with a granularity not even imagined twenty years ago.

In particular, mobile phone datasets provide a very rich view into the social interactions and the physical movements of large segments of a population. The voice calls and text messages exchanged between people, together with the call locations (recorded through cell tower usages), allow us to construct a rich social graph which can give us interesting insights on the users' social fabric, detailing not only particular social relationships and traits, but also regular patterns of behavior both in space and time, such as their daily and weekly mobility patterns [5, 8, 10].

Demographic factors play an important role in the constitution and preservation of social links. In particular concerning their age, individuals have a tendency to establish links with others of similar age. This phenomenon is called age homophily [7], and has been verified in mobile phone communications graph [2, 9] as well as the Facebook graph [11].

Economic factors are also believed to have a determining role in both the social network's structure and dynamics. However, there are still very few large-scale quantitative analyses on the interplay between economic status of individuals and their social network. In [6], the authors analyze the correlations between mobile phone data and banking transaction information, revealing the existence of social stratification. They also show the presence of socioeconomic homophily among the networks participants using users' income, purchasing power and debt as indicators.

In this work, we leverage the socioeconomic homophily present in the cellular phone network to generate inferences of socioeconomic status in the communication graph. To this aim we will use the following data sources: (i) the Call Detail Records (CDRs) from the operator allow us to construct a social graph and to establish social affinities among users; (ii) banking reported income for a subset of their clients obtained from a large bank data source. We then construct an inferential algorithm that allows us to predict the socioeconomic status of users close to those for which we have banking information. To our knowledge, this is the first time both mobile phone and banking information has been integrated in this way to make inferences based on a social telecommunication graph.

2 Data Sources

2.1 Mobile Phone Data Source

The data used in this study consist of a set $\mathcal{P}$ of *Call Detail Records* (CDRs), composed of voice calls and text messages from a Mexican telecommunication company (*telco*) for a 3 month period.

Every CDR $p \in \mathcal{P}$ contains the phone numbers of the caller and callee $\langle p_o, p_d \rangle$, which are anonymized using a cryptographic hash function for privacy reasons, the starting time p_t, and, in the case of voice calls, the call duration p_s. The latitude and longitude of the antenna used for each call $\langle p_y, p_x \rangle$ are also given for a subset of the data.

Given that our collection $\mathcal{P}$ of CDRs are coming from one telephone company, we are able to reconstruct all communication links between clients of this company, as well as communications between the clients and other users, but we have no information on communications where neither users are clients of our telco company.

If we define N as the set of clients of the telco, and $\mathcal{P}_N \subseteq \mathcal{P}$ as the calls where $\forall p \in \mathcal{P}_N,\ p_o \in N \wedge p_d \in N$, we create a communications graph $\mathcal{G}_N$ which contains only the users from the telco, and all the calls exchanged between them.

2.2 Banking Information

For this study we also used account balances for over 10 million clients of a bank in Mexico for a period of 6 months, denoted $\mathcal{B}$. The data for each client $b \in \mathcal{B}$ contains his phone number b_p, anonymized with the same hash function used in $\mathcal{P}$, and the reported income of this person over 6 months $b_{s_0}, \ldots, b_{s_5}$. We average these 6 values to get b_s, an estimate of a user's income.

The bank also provided us demographic information for a subset of its clients $\mathcal{A} \subseteq \mathcal{B}$. For each user $u \in \mathcal{A}$, we are given the age u_a of the user, which allows us to observe differences in the income distribution according to the age. In another line of work, homophily with respect to age has been observed and used to generate inferences [3].

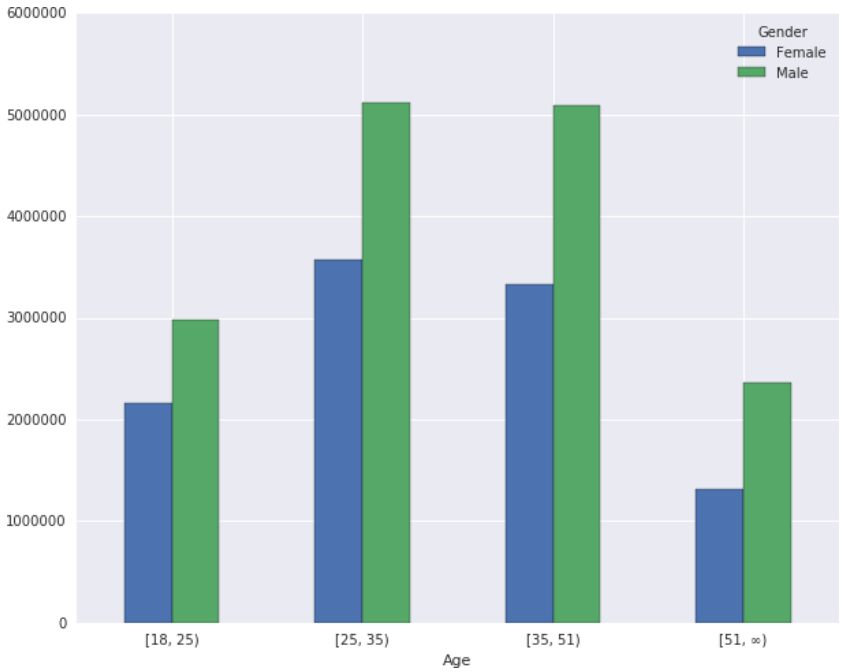

Fig. 1. Amount of users in $\mathcal{B}$ by gender and age.

Figure 1 shows the distribution of users in $\mathcal{B}$, according to their age range and gender. Figure 2 shows the distribution of income, according to the age range (generated by taking 5 years intervals for the age). It is interesting to note how the median income increases with the age, up to the 60–65 years range (the retirement age in Mexico). After 65 years old, the median income rapidly decreases.

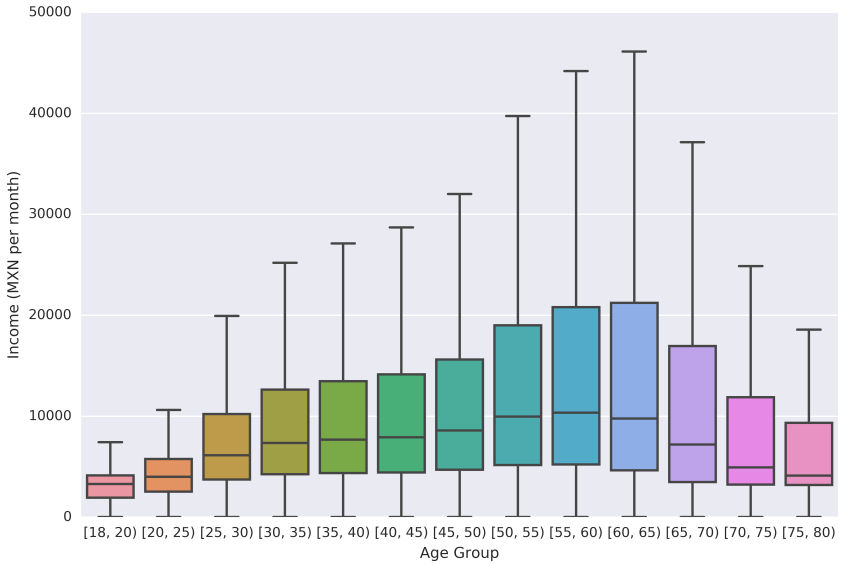

Fig. 2. Distribution of income b_s as a factor of age b_a. This is consistent with data from median house income in Mexico [4].

2.3 Bank and Telco Matching

Since the phone numbers in each call p_o and p_d are anonymized with the same hash function as the phone number in the bank data, b_p, we can match users to their unique phone to create the social graph:

$$G = \mathcal{P} \bowtie_{p_o=b_p} \mathcal{B} \bowtie_{p_d=b_p} \mathcal{B}$$

where $\bowtie$ denotes the inner join operator. G includes income information for the subset of the social graph that appears in the bank data, so $\forall g \in G$ we have its phone number g_p, its average income over 6 months g_s, and its age g_a. This graph has a total of 2,027,554 nodes with 5,044,976 edges, which represent 29,599,762 calls and 5,476,783 text messages.

2.4 Outlier Filtering

The dataset contains information about bank and telco users, some of which may not directly correspond to a human user, or may not have useful information for our research. Most of the telco users in the first case are already filtered by the intersection (INNER JOIN). To make sure the users are relevant enough for this study, we only keep the users which have:

– More than 5 calls in either direction.

– A monthly income of at least $1000. The value is expressed in Mexican pesos (MXN)[3].
– A monthly income in the 99th percentile (i.e. we filter users with a monthly income in the top 1%).

2.5 Unequal Distribution of Income

We provide here some observations of the distribution of income of the bank clients. These observations correspond to the filtered dataset, obtained after applying the filters of the previous section.

Figure 3 shows the Lorenz curve, graphical representation of the distribution of income. The X axis plots the cumulative share of clients, ordered from lowest to highest incomes. The Y axis plots the fraction of the total income that they have.

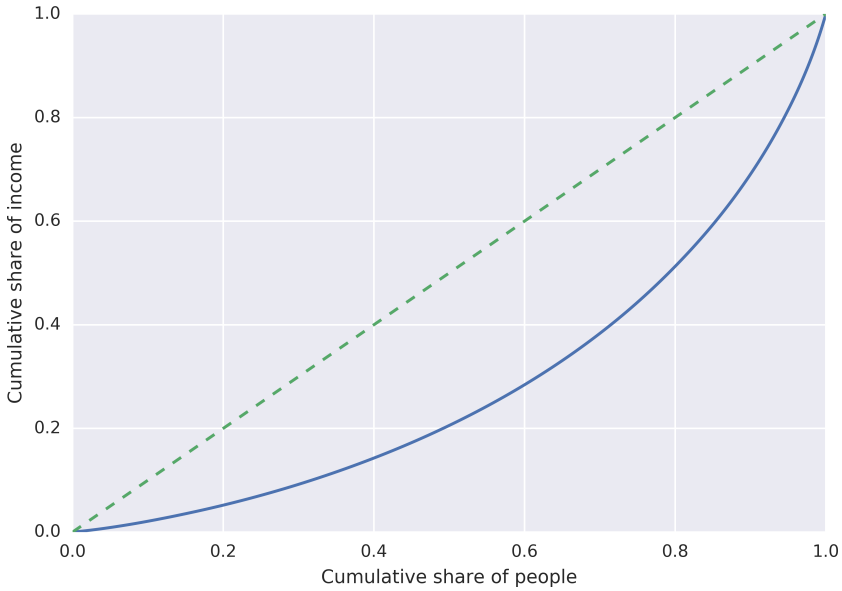

Fig. 3. Lorenz curve representing the distribution of income of bank clients.

From the Lorenz curve, we can compute the Gini coefficient, as the area that lies between the line of perfect equality (the line at 45°) and the Lorenz curve over the total area under the line of equality. The Gini coefficient obtained is $G = 0.45$. According to the World Bank [1], the Gini coefficient for the whole population

[3] At the time of writing (July 14, 2016), 1000 Mexican pesos are equivalent to 54 US dollars.

of Mexico was 0.481 in 2012. Our result is consistent with this information, since the income inequality is expected to be lower within the bank clients than within the whole population of the country.

Looking at the cumulative share of the clients with highest incomes, we observe that the top 10% of clients accumulate 33% of the total income; the top 20% accumulate 50.5%; and the top 30% accumulate 63.1% of the total income.

3 Inference Methodology

3.1 Income Homophily

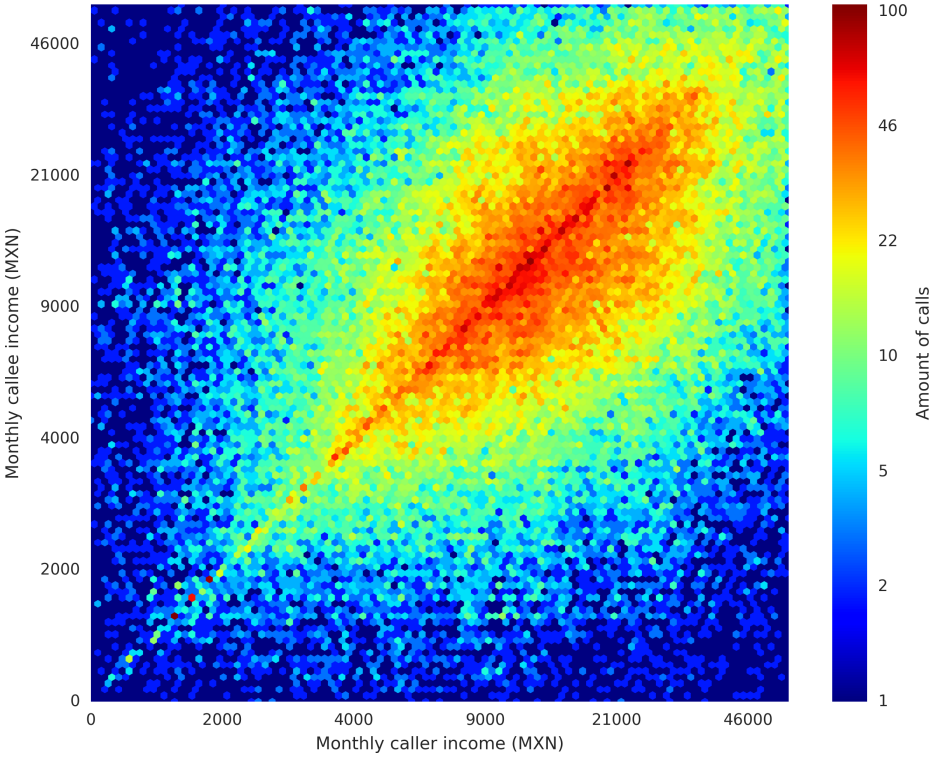

Fig. 4. Heatmap showing the number of calls between users, according to their monthly income. There is a higher probability that the callee and the caller have similar income levels.

The main contribution of this work is the estimation of the income of the telco users for which we lack banking data, but have bank clients in their neighborhood of the network graph. To show the feasibility of this task, we first show the

existence of a strong income homophily in the telco graph as is evidenced in Figure 4.

For each pair $\langle o, d \rangle \in G$, we define X as the set of incomes for callers and Y as the set of incomes for callees. According to what we can observe in Figure 4, X and Y should be significantly correlated. Given the broad non-Gaussian distribution of the income's values, we choose to use a rank-based measure of correlation which is robust to outliers. Namely we computed the *Spearman's rank correlation* to test the statistical dependence of sets X and Y:

$$r_s = \rho_{\mathrm{rank}(X)\,\mathrm{rank}(Y)} = \frac{\mathrm{cov}(\mathrm{rank}(x), \mathrm{rank}(y))}{\sigma_{\mathrm{rank}(X)}\sigma_{\mathrm{rank}(Y)}} \tag{1}$$

this coefficient gives us a correlation coefficient of $\mathbf{r_s = 0.474}$. We also compared our result with a randomized null hypothesis, where links between users are selected randomly disregarding income data, obtaining a p-value of $p < 10^{-6}$. These values for r_s and p show a strong indication of income homophily among users in our communication graph. This observation is consistent with the results reported in [6].

We can take advantage of this homophily to propagate income information to the rest of our graph $\mathcal{P}$, where we don't know the income of all the users.

3.2 Prediction Algorithm

Instead of predicting the exact value of a user's income, our strategy is to distinguish between only two income categories, $R_1 = [1000, 6300)$ and $R_2 = [6300, \infty)$, that is, users with low or high income respectively, which we place into two distinct groups $H_1, H_2 \subseteq G$ depending on g_s, the users' income:

$$g \in H_i \iff g_s \in R_i$$

We define the set Q as the group of users having at least one connection link to bank clients. For each user $q^j \in Q$, we compute the number of outgoing calls a_i^j to the category H_i. Our hypothesis, given the observed homophily, is that if a user q^j has a higher number of calls a_i^j to the category H_i than the other category, it would be more likely to belong to the H_i income category. In other words, a person is usually in the same income category as the majority of people it calls.

A straightforward approach would be to define the income category of a user as the category where most of its contacts belong. The problem with this approach is that it does not factor in the higher uncertainty in our estimates for users with fewer calls. To address this uncertainty, instead of using calling frequencies to define the probability of a user belonging to the high income category, we use the amount of calls a_i^j as parameters defining a Beta distribution for the probability of belonging to a given category. We have therefore taken a Bayesian rather than a frequentist approach to income prediction.

We define B^j as the Beta probability distribution function for each user:

$$B^j\left(x;\alpha^j,\beta^j\right)=\frac{1}{B\left(\alpha^j,\beta^j\right)}x^{\alpha^j-1}\cdot(1-x)^{\beta^j-1} \tag{2}$$

where $\alpha^j=a_1^j+1$ and $\beta^j=a_2^j+1$ are the parameters of the Beta distribution, and B is the beta function, defined as:

$$B\left(\alpha,\beta\right)=\frac{\Gamma(\alpha)\cdot\Gamma(\beta)}{\Gamma(\alpha+\beta)}. \tag{3}$$

Note that the above equation defines a distinct distribution for each user. Having obtained the Beta distribution for the probability of belonging to the high income category, we then find the lowest 5 percentile p_{lower} for this probability. If p_{lower} is above a given threshold τ, we set the user's income to H_2, otherwise we set his income category to H_1. We note that this criteria takes into account both the mean and the broadness (uncertainty) of the distribution. We also note that the category assigned to a user depends not only on its Beta distribution but also on our choice of τ.

4 Results

4.1 Evaluation of Performance

We describe in this section the validation of our methodology. We examine the true positive (TPR) and false positive (FPR) rates, TPR = TP / P and FPR = FP / N, where TP is the number of correctly predicted users with high income, P is the total number of users with high income, FP is the number of users incorrectly classified as having high income, and N is the total number of users with low income.

In Figure 5 we plot the ROC (*Receiver Operating Characteristic*) curve, showing TPR and FPR for the set of possible values of τ. We see that our methodology clearly outperforms random guessing (dashed straight line). We can summarize our performance by calculating the AUC (*Area Under the Curve*) which in Figure 5 is AUC = 0.74. Note that random guessing would give a value of AUC $\simeq$ 0.50.

Alternatively we can evaluate the performance of our model by computing its accuracy for a given threshold τ of FPR. Figure 6 shows the accuracy obtained as a function of FPR, where the accuracy is computed as the ratio between TP and TN over the total population for a certain FPR. The best accuracy obtained is 0.71 for $\tau = 0.4$.

4.2 Comparison with Other Inference Methods

We applied two other inference methods to the same data and compared their accuracies to our Bayesian model.

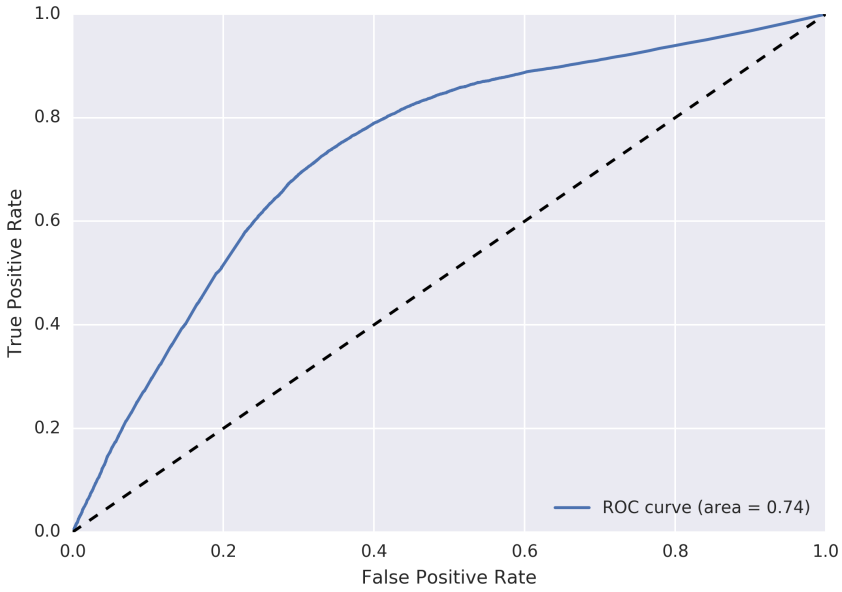

Fig. 5. ROC curve for prediction procedure. We observed an AUC = 0.74 indicating that our predictor is better than a random predictor (AUC $\simeq$ 0.50).

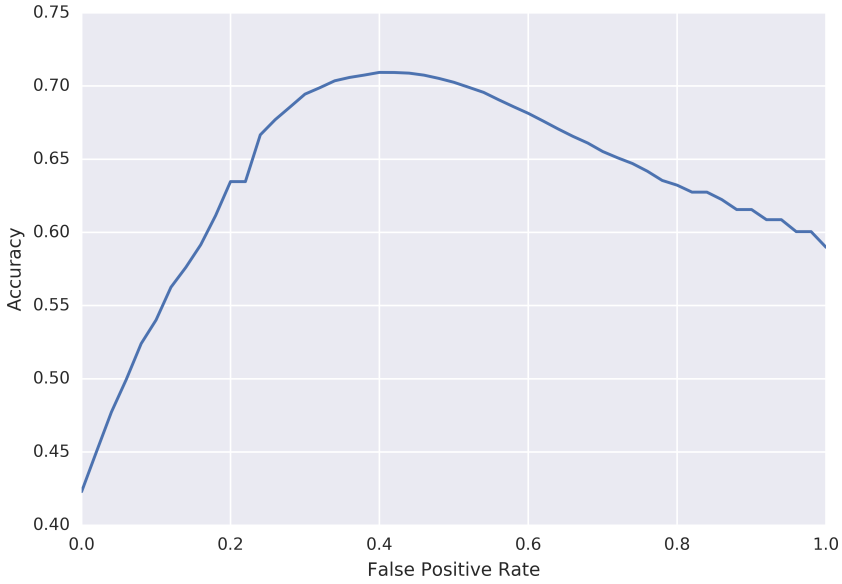

Fig. 6. Accuracy as a function of FPR. The best accuracy obtained is 0.71.

- **Random selection** which chooses randomly the category for each user.
- **Majority voting** which decides whether a user is in the high or low income category depending on the category of the majority of its contacts. In case of a tie, the category is chosen randomly.

The accuracy of the first method is as expected 0.5, while the accuracy for majority voting is 0.66. With the Bayesian method we obtain an accuracy of 0.71.

5 Extension to Multiple Categories

We present here how the methodology described in Section 3 for two categories can be extended to multiple categories. To this end, we separate the income values into five distinct groups $H_1, \ldots, H_5 \subseteq G$ of increasing wealth where:

$$g \in H_i \iff g_s \in R_i$$

and the income ranges are set as follows (in Mexican pesos):

$$\begin{aligned} R_1 &= [1000, 2500) \\ R_2 &= [2500, 7500) \\ R_3 &= [7500, 20000) \\ R_4 &= [20000, 50000) \\ R_5 &= [50000, \infty) \,. \end{aligned}$$

Again, we define the set Q as the group of users having at least one connection link to bank clients. For each user $q^j \in Q$, we compute the number of outgoing calls a_i^j to the category H_i. We use the amount of calls a_i^j as parameters defining a Dirichlet distribution for the probability of belonging to a given category. We define below the Dirichlet probability distribution function D^j:

$$D^j\left(x_1, \ldots, x_5; \alpha_1^j, \ldots, \alpha_5^j\right) = \frac{1}{B\left(\alpha\right)} \prod_{i=1}^{5} x_i^{\alpha_i^j - 1} \tag{4}$$

where $\alpha_i^j = a_i^j + 1$ are the parameters of the Dirichlet distribution, and B is the multivariate beta distribution function, defined by:

$$B\left(\alpha_1, \ldots, \alpha_k\right) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)} \tag{5}$$

Note that the above equation defines a distinct Dirichlet distribution for each user. For each of these distributions, we computed the marginal probability functions across all different categories, which result in Beta distributed functions,

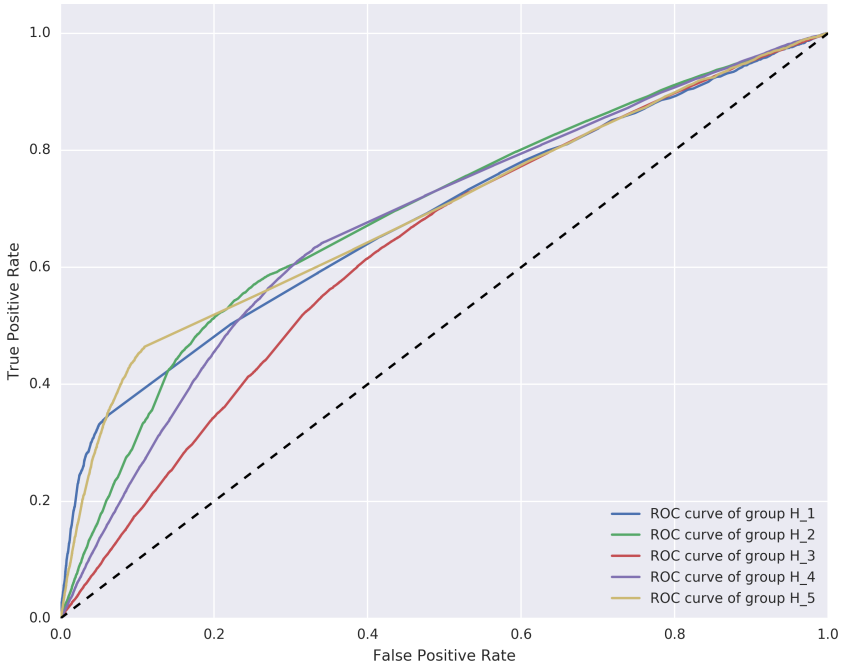

Fig. 7. ROC curves for multiclass problem. The performances observed are: $\mathrm{AUC}_1 = 0.68$, $\mathrm{AUC}_2 = 0.69$, $\mathrm{AUC}_3 = 0.63$, $\mathrm{AUC}_4 = 0.68$, $\mathrm{AUC}_5 = 0.69$. These predictors perform better than the random case, and have a similar performance (with exception of category 3).

and use them to get the lowest 5 percentiles (p^i_{lower}) in each case $i = 1, \ldots, 5$ which can be compared to assign a category to each user.

In order to gain an intuition on how the classification extends to the multiple category case, we constructed for each category i a binary classifier by using the computed p^i_{lower} score and a given threshold τ. In each case we sweep the threshold τ and compute the resulting ROC curves as shown in Figure 7.

We observed the performance for the different categories: $\mathrm{AUC}_1 = 0.68$, $\mathrm{AUC}_2 = 0.69$, $\mathrm{AUC}_3 = 0.63$, $\mathrm{AUC}_4 = 0.68$, $\mathrm{AUC}_5 = 0.69$. In all cases, the predictor performs better than the random case.

6 Conclusion

This work is based on the combination of two data sources of mobile phone records and banking information. We showed that there is a significant level of homophily between the income of the participants of a call, and based on this property, we presented a Bayesian approach to infer the income category of users in the graph for which we don't have banking data.

We first classified users into 2 categories depending on their income. To this end, we computed the number of calls each user u makes to members of the same and different categories, and we constructed a Beta distribution for the probability of user u belonging to each category. We later validated this approach by constructing the ROC curve and computing its accuracy, and compared it to random guessing and to a simpler method based on majority voting. We were able to validate that the method presented outperforms the other two. Finally, we showed that this approach can be extended to more than 2 categories by using a Dirichlet distribution.

Our proposed inference methodology is useful in concrete applications, since it provides an estimation of socio-economic attributes of users lacking banking history, based on their communication network. We also note that this methodology is not restricted to the inference of socio-economic attributes, but is equally applicable to any attribute that exhibits significant homophily in the network.

References

1. World Bank. World bank open data, 2016. [Online, accessed 14-July-2016].
2. Joshua Blumenstock and Nathan Eagle. Mobile divides: gender, socioeconomic status, and mobile phone use in Rwanda. In *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development*, page 6. ACM, 2010.
3. Jorge Brea, Javier Burroni, Minnoni Martin, and Carlos Sarraute. Harnessing mobile phone social network topology to infer users demographic attributes. In *ACM SIGKDD*. ACM, 2014.
4. Gallup. Worldwide median house income, 2013. [Online, accessed 16-may-2016].
5. Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.
6. Yannick Leo, Eric Fleury, Carlos Sarraute, José Ignacio Alvarez-Hamelin, and Màrton Karsai. Socioeconomic correlations in communication networks. In *Fourth conference on the Analysis of Mobile Phone Datasets (NetMob 2015)*, 2015.
7. Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.
8. Nicolas Ponieman, Alejo Salles, and Carlos Sarraute. Human mobility and predictability enriched by social phenomena information. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 1331–1336. ACM, 2013.
9. Carlos Sarraute, Pablo Blanc, and Javier Burroni. A study of age and gender seen through mobile phone usage patterns in Mexico. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 836–843. IEEE, 2014.
10. Carlos Sarraute, Carolina Lang, Nicolas B Ponieman, and Sebastian Anapolsky. The city pulse of Buenos Aires. In *Workshop Big Data & Environment*, 2015.
11. Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. The anatomy of the Facebook social graph. *Structure*, 5:6, 2011.