Proc. Conference on Data Mining and Data Warehouses (SiKDD 2017), pp. 1-4
October 9, 2017, Ljubljana, Slovenia

Building Client's Risk Profile Based on Call Detail Records

Zala Herga[1, 3], Casey Doyle[2], Stephen Dipple[2], Caleb Nasman[2], Gyorgy Korniss[2], Boleslaw Szymanski[2], Janez Brank[1], Jan Rupnik[1], and Dunja Mladenić[1, 3]

[1]Jožef Stefan Institute, Jamova 39, Ljubljana, Slovenia

[2]Rensselaer Polytechnic Institute, 110 8th Street,Troy, NY 12180, USA

[3]Jožef Stefan Postgraduate School, Jamova 39, Ljubljana, Slovenia

{*zala.herga,janez.brank,jan.rupnik, dunja.mladenic*}*@ijs.si*

{*doylec3, korniss, nasmancc1, dippls, szymab*}*@rpi.edu*

ABSTRACT

Data collected from mobile phones can be used to uncover underlying social network dynamics and individual's behavioral patterns. Based on a Call Details Records dataset, we build a weighted, directed network and analyze it's properties. In addition to node-level network measures we extract an extensive consumption and mobility-based feature set. We show that extracted network and consumption features can be used to model individual's risk profile.

Keywords

Mobile Phone Network, CDR, Supervised learning

1. INTRODUCTION

The Call Detail Records (CDR) dataset is a relatively standard dataset obtained by mobile phone operators. One record in the CDR dataset corresponds to a communication event between two mobile phone users and includes time stamp, type of event (call, text), direction (in- or outgoing) etc. This data type reveals behavioral patterns that can be used to identify user's personality [2], spending habits [11] or socioeconomic level [5]. Here, we are interested in using the data to build each client's risk profile; in particular, we attempt to use this data to predict user defaults. To this end, we focus our analysis around whether the clients phone number was blocked at the end of month (indicating issues potentially related to the defaulting behavior), using this data to label clients as good or defaulted. The dataset used is completely anonymised.

Structure of the rest of the paper is as follows: Section 2 presents characteristics of the network built from the CDR, Section 3 describes feature extraction, Section 4 presents probability of default models and their evaluation and Section 5 concludes the paper.

2. NETWORK PROPERTIES

As the first step in analyzing the dataset and gaining an understanding of how the users operate we define the structure of the network. Here, we treat the mobile data set as a social network where each node is an individual and each edge represents a connection between them and another individual. Wherever possible, we use weighted, directed edges to preserve the strength of the connection between individuals [4]. Weights are assigned based on the frequency of outgoing communications between the source node and the target node. Where applicable, we use this metric to define the distance between two nodes as $w_{avg}/w_{i\to j}$ where w_{avg} is the average weight of all connections in the network and $w_{i\to j}$ is the weight of the connection between the source (i) and the target (j) [6]. Wherever it is not feasible to use a weighted edge scheme, we create an unweighted graph using a cutoff to define how many outgoing communications from one node to another constitutes a connection (ie we use the frequency to define whether a connection exists at all, and all connections are still directed but have equal weight) [5]. Using a low cutoff introduces a lot of noise into the system and is less representative of a true social network as many of the edges are too weak to accurately indicate a social connection between two individuals, but choosing too high of a cutoff restricts the network and discards potentially valuable data connecting nodes and communities together.

Using these methods, our data set translates to a network with a giant component comprising 99.14% of the it. Of course, the number of edges and size of the giant component decreases quickly when the unweighted cutoff scheme is used (Fig. 1). The size of the giant component decreases linearly with increases in the cutoff, while the decrease in the number of edges levels off as a power law with $\gamma \approx 0.75$. The degree distribution also changes slightly with the cutoff than without; in both cases the distribution has a fat tail that is well approximated by a power law, but the exponent increases with a larger cutoff. For the general case of the weighted edges with no cutoff, the power law tail has an exponent of $\gamma \approx -4.3$ while with a high cutoff such as thirty the power law tail exhibits an exponent of $\gamma \approx -6$ (Fig. 2), both in general agreement with prior work on mobile network data [5, 7]. Similarly, the distribution of node strengths (defined as the sum of the weights of its adjacent edges) also exhibits a heavy tailed decay as expected [7].

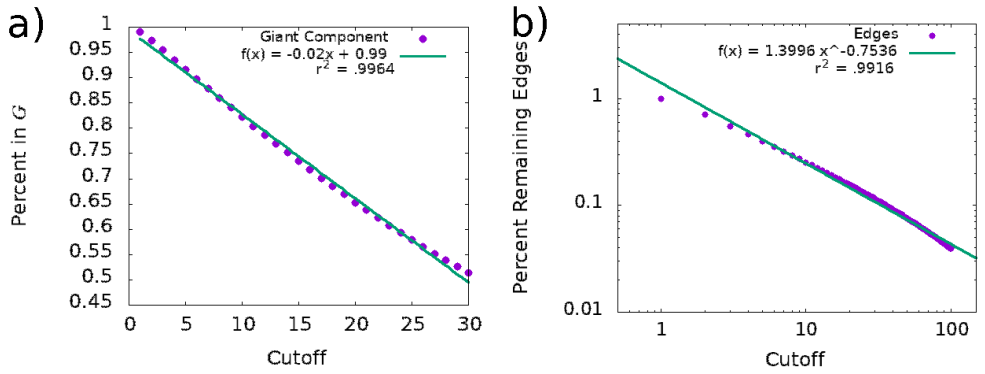

Figure 1: (a) The size of the giant component G as it decreases linearly with higher cutoff criteria to form an edge between two nodes. (b) The total number of edges in the system decreases as a power law with $\gamma \approx 0.75$.

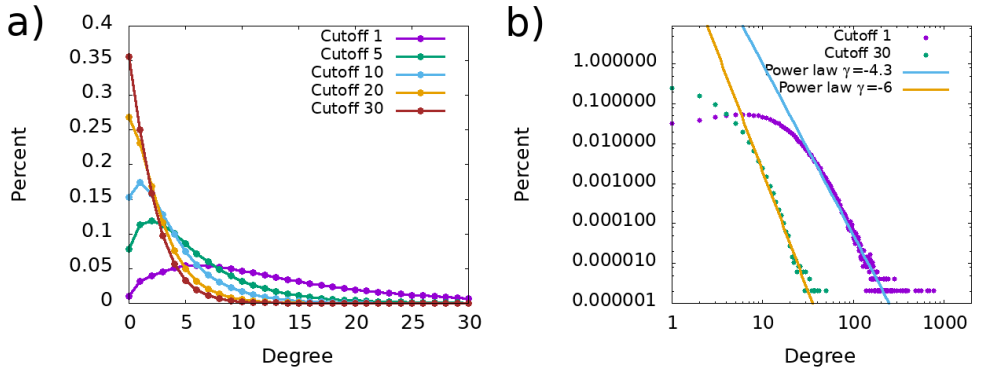

Figure 2: (a) The degree distribution for various cutoff criteria to create an edge between two nodes. As the cutoff increases, the peak of the distribution shifts left until it peaks at zero. (b) The same distributions on a log-log scale to highlight the power law tail of the distributions. Shown here are the two extreme cutoffs tested in order to highlight the increase in the gamma value for higher cutoffs.

Additionally, we study the distribution of some higher level node based measures such as reciprocity[14], which is surprisingly low. Without a cutoff only 38.65% of all links are reciprocated, while higher cutoffs increase the fraction of reciprocated links up to a maximum of only 41.57% when the cutoff is fifteen. We further measure the reciprocity using the weighted network scheme by defining the weighted reciprocity[12, 13] as $R_{ij} = |w_{ij} - w_{ji}|/(w_{ij} + w_{ji})$. Using this metric, the network shows an average weighted reciprocity of only .3235, further indicating the low reciprocity of the network.

Finally, we use node centrality to measure how the nodes position themselves in within the communication paths across the network (nodes with high centrality are most likely to connect communities and therefore are very important to the study of how risk patterns propagate across the network). For this purpose, we utilize the closeness centrality[1, 10, 3], a node level measurement that utilizes the shortest paths across the network to identify where nodes lie in the network structure. Specifically, it is a ranking of the distance from the node in question to every other node, defined as $C_C(i) = (N-1)(\sum_{i \neq j} d_{ij})^{-1}$ where C_C is the closeness centrality and d_{ij} is the length of the shortest path between nodes i and j (assuming a path exists). Unfortunately this measure only works on connected graphs, so to analyze the full unconnected graph, we also study the harmonic closeness centrality[9, 8], defined instead as $C_H(i) = \frac{1}{n-1} \sum_{j \neq i} \frac{1}{d_{ij}}$. As seen in Fig. 3, the closeness centrality has a tightly grouped, high density of relatively high values. This implies a very well connected graph such that the shortest path between any two nodes is low. This can further be seen via the harmonic centrality, which also has a relatively low density of low scoring individuals even with the inclusion of nodes not within the giant component. This implies that even the nodes that are not connected to the giant component tend to form small, tightly connected communities of their own.

3. FEATURE EXTRACTION

After understanding the network dynamics, our aim was to build individual's behavioral patterns. For that reason we extracted from the CDR three types of behavior-related features: individual's consumption, social network and mobility. Some of the features were extracted for different time window (e.g. per day, per week, hour of day), separately for incoming and outgoing events and/or separately for event type (call, text). We also added another more technical category which relates to individual's position in the underlying network. Together, more than 6000 features were extracted. Each category is described in more detail below.

1. Consumption features: These features are related to individual's usage of the mobile phone. We extracted for each individual the number of all calls, number of all texts, total duration of calls, average duration of calls, average time between consecutive events.

2. Social network features: This type of feature focuses

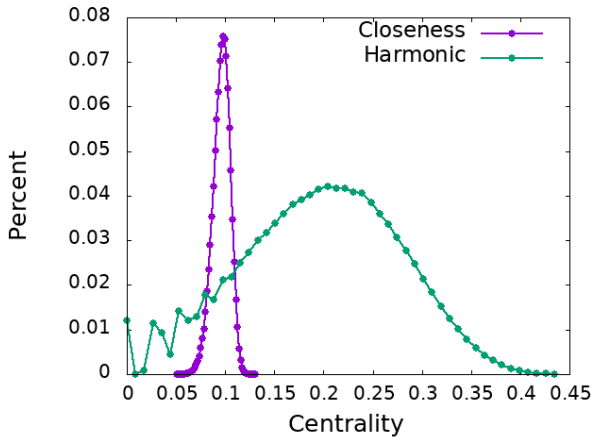

Figure 3: Distribution of closeness and harmonic centrality scores across the network where higher scores indicate a more central position in the network. The closeness centrality only considers nodes in the giant component (and the distribution is therefore only calculated for those nodes). The harmonic centrality includes all nodes in the network.

on the number of contacts and reciprocated events: the number of unique contacts, the number of contacts with which individual exchanges on average at least 5 texts per week / 2 calls per week, the number of reciprocated call events, the median time between reciprocated call events, and the median time to answer text.

3. Mobility features: These features that are based on used BTS tower location and include the average daily radius of gyration, the average distance traveled per day of week, the popular cell towers that sum up to 90% of records, and the average number of unique cell towers used per week.

4. Node level network measures: These features all rely on the individuals location within the social network built off of their usage statistics. The details of these metrics are discussed in Section 2, and represent each nodes level of importance to the overall social network as well as how deeply embedded the individual is.

3.1 Geographic analysis

Geographic analysis was performed to help us with the specific goal of building individual's risk profile. Analyzing geographic features requires a definition of their location that considers that most people connect to many different cell towers over the course of a three month period. For our purposes, we use each user's top two most used cell towers. We assign an individual to both their most used and second most used towers to account for the likelihood that a user will spend large amounts of time both at their residence and their workplace. From there we analyze the number of people that exhibit default behavior for each tower or district and identify high risk geographic regions. Based on that analysis we calculated empirical probability of default for *each cell tower*. These probabilities were used as two additional features, one for each of the two most commonly used towers of each user.

4. MODELING

Our aim was to model probability of default for each client based on extracted phone usage patterns and node-level network measures. We present the results of fitting several linear regression models with varying parameters. Features that are described in previous sections were used for modeling, and all features were normalized to standard score (z-score). We divided our dataset into train- and test set in 70:30 ratio.

We started with a linear model (labeled as *glm-6* in Figure 4) that was based only on six predictor variables: *frequency* (corresponds to the node's strength), *duration* (of user's call events; sum), *degree*, *harmonic centrality* and the two *geographic* (cell tower PD) variables. We chose with these features because we believed that they carry a lot stronger signals in contrast to the other 6048 features. The p-value is < 0.01 for all, except for *duration*, which has a p-value of 0.97. Overall this implies that network measures are good predictors for default behavior.

4.1 Principal Component Analysis

Further, when dealing with larger amount of features (6048) principal component analysis (PCA) was performed on the train set for feature space reduction. PCA is a method that decomposes the feature space into principal components (eigenvectors) and also provides information about how much variance in the data each component explains. Selection of a subset of PCA components reflects a trade-off between 1) model simplicity (we want to include a moderate number of features in our models) and 2) total variance explained by the component subset. All features were subjects to PCA, except for the 6 features that were used in *glm-6* model described above. Those were added to models in their original (but standardized) form.

We ordered the obtained PCA components decreasingly by explained variance of the data. The first component explains 20% of the variance, the second 7%, the first ten components together 37%, first thirty together 42%, and first five hundred sum up to 66%. We then created two linear models based on reduced feature subsets: first, using 30 PCA components and second, using 500 components (*pca-30* and *pca-500* in Figure 4). Because many variables in *pca-500* have large p-values, we fitted another model that didn't include those variables with p-value ≥ 0.5 (*pval-05*).

4.2 Oversampling

Only about 0.25% of users in the underlying dataset exhibited default behavior, which makes the dataset very unbalanced. For that reason, we implemented a simple oversampling method on train set: we multiplied defaulted users (and their features) by 20. The model using this method, *oversampled-20*, is also presented in Figure 4. Surprisingly, the oversampled dataset not only does not improve performance, but can be seen to provide slightly worse results than the originally unbalanced dataset.

4.3 Evaluation

Model comparison is presented in Table 1. We can see that at 95%, the level recall is high, up to 0.91 for both models based on 500 PCA components. Precision is low for all models due to the unbalanced dataset, but even with that drawback, our models still perform far better than random models.

Model	random	glm-6	pca-30	pca-500	pval-05	oversampled-60
Recall	0.05	0.13	0.79	0.90	0.91	0.86
Precision	0.003	0.007	0.042	0.049	0.049	0.046

Table 1: Recall and precision at 95% level for each of the models presented in Figure 4.

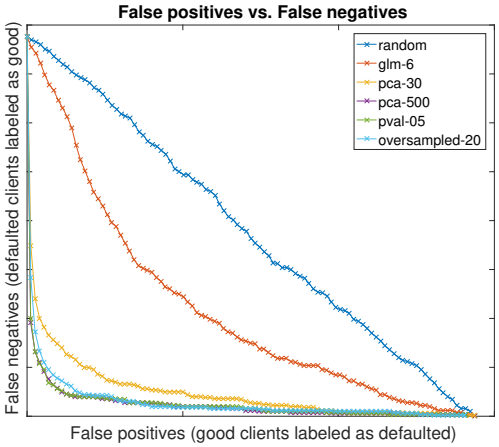

Figure 4: This graph presents prediction results on test set of the fitted models. y-axis corresponds to the false negatives (clients, that we're labeled as *good* but really defaulted), while x-axis corresponds to false positives. Results are shown for probability thresholds $0 - 1$ with step 0.01. *pca-500* (purple) is to great extent covered by *pval-05* (green) since both models provide very similar results.

5. CONCLUSION

This paper presents an analysis of a mobile phone data using a social network representation and various prediction models to understand default patterns. The analysis on the underlying network reveals a large giant component such that most nodes have at least some path to any other node in the network. Further, both the nodes within and without the giant component exhibit relatively high centrality scores; meaning that nodes are form tightly connected communities such that the path between nodes is generally quite short. Further, many nodes have a high degree and the degree distribution exhibits a heavy power law-like tail. Using many of these properties as features, we were able to make even more accurate predictive models of default.

Our model evaluation shows that there are many variables that carry weak signals about user behavioral patterns that have a strong predictive power when aggregated together. The unbalanced nature of the dataset makes the fitted models have a high recall but low precision, yet they strongly outperform the random model in both measures.

There is still a lot of space for improvement in the modeling including testing more complex oversampling methods, fitting additional models (SVM, LASSO, ANN), and including additional node-level network measures and community detection analysis.

6. ACKNOWLEDGEMENTS

This work was supported by RENOIR EU H2020 project under Marie Skodowska-Curie Grant Agreement No. 691152 as well as in part by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053 (the Network Science CTA), by the Office of Naval Research (ONR) Grant No. N00014-15-1-2640, and by NSF Grant No. DMR-1560266 under the REU Program. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies either expressed or implied of the Army Research Laboratory or the U.S. Government.

7. REFERENCES

[1] A. Bavelas. Communication Patterns in TaskOriented Groups. *J. of the Acoustical Society of America*, 22(6):725–730, nov 1950.

[2] Y.-A. de Montjoye, J. Quoidbach, F. Robic, and A. Pentland. Predicting personality using novel mobile phone-based metrics. In *SBP*, pages 48–55, 2013.

[3] L. C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239, jan 1978.

[4] M. S. Granovetter. The Strength of Weak Ties. *American J. of Sociology*, 78(6):1360–1380, May 1973.

[5] S. Luo, F. Morone, C. Sarraute, M. Travizano, and H. A. Makse. Inferring personal economic status from social network location. *Nature communications*, 8:15227, may 2017.

[6] M. E. J. Newman. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E*, 64(1):016132, Jun 2001.

[7] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási. Structure and tie strengths in mobile communication networks. *PNAS*, 104(18):7332–6, May 2007.

[8] T. Opsahl, F. Agneessens, and J. Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32(3):245–251, jul 2010.

[9] Y. Rochat. Closeness centrality extended to unconnected graphs: The harmonic centrality index. In *ASNA*, number EPFL-CONF-200525, 2009.

[10] G. Sabidussi. The centrality index of a graph. *Psychometrika*, 31(4):581–603, dec 1966.

[11] V. K. Singh, L. Freeman, B. Lepri, and A. S. Pentland. Predicting spending behavior using socio-mobile features. In *SocialCom*, pages 174–179. IEEE, 2013.

[12] T. Squartini, F. Picciolo, F. Ruzzenenti, and D. Garlaschelli. Reciprocity of weighted networks. *Scientific Reports*, 3(1):2729, dec 2013.

[13] C. Wang, A. Strathman, O. Lizardo, D. Hachen, Z. Toroczkai, and N. V. Chawla. Weighted reciprocity in human communication networks. aug 2011.

[14] S. Wasserman and K. Faust. *Social network analysis : methods and applications*. Camb. Univ. Press, 1994.