Information Sciences, **500**:271-228, May30, 2019

Predicting complex user behavior from CDR based social networks

Casey Doyle[a], Zala Herga[b,c], Stephen Dipple[a], Boleslaw K. Szymanski[a,d,*], Gyorgy Korniss[a], Dunja Mladenić[b,c]

[a] *Rensselaer Polytechnic Institute, 110 8th Street,Troy, NY 12180, USA*
[b] *Jožef Stefan Institute, Jamova 39, Ljubljana, Slovenia*
[c] *Jožef Stefan International Postgraduate School, Jamova 39, Ljubljana, Slovenia*
[d] *Społeczna Akademia Nauk, Łódź, Poland*

Abstract

Call Detail Record (CDR) datasets provide enough information about personal interactions of cell phone service customers to enable building detailed social networks. We take one such dataset and create a realistic social network to predict which customer will default on payments for the phone services, a complex behavior combining social, economic, and legal considerations. After extracting a large feature set from this network, we find that each feature poorly correlates with the default status. Hence, we develop a sophisticated model to enable reliable predictions. Our main contribution is a methodology for building complex behavior models from very large sets of diverse features and using different methods to choose those features that perform best for the final model. This approach enables us to identify the most efficient features for our problem which, unexpectedly, are based on the number of unique users with whom the given user communicates around the Christmas and New Year's Eve holidays. In general, features based on the number of close ties maintained by a user perform better than others. Our resulting models significantly outperform the methods currently published in the literature. The paper contributes also a systematic analysis of properties of the network derived from CDR.

Keywords: social networks; complex behavior prediction; probability of

*Corresponding author
Email address: szymab@rpi.edu (Boleslaw K. Szymanski)

default; feature selection; Call Detail Record dataset

1. Introduction

Call Detail Record (CDR) datasets, created from cell phone logs of large groups of people, have become common in studying human behavior thanks to the large amount of detailed data they provide [5]. These datasets typically include both basic information about the users (age, gender, location) and records of calls and text messages including the time, location, and direction of each communication.

Here, we analyze such a dataset and focus on the properties of the underlying social network that can be obtained from the data. First, we discuss various methods for building the network, aiming to mitigate the noise inherent within the cell phone records and address other issues identified in prior work such as the definitions of links, communities, reciprocity, and data types suitable for the study. These considerations are particularly relevant to our application due to the large amounts of noise and potential biases inherent to various network representation schemes [12, 15, 17, 14, 28]. In this paper, we combine techniques based on geographic and usage features along with higher level social network measures such as centrality and community structure to predict the probability of customers defaulting on their accounts.

The specific CDR dataset used here includes the detailed call history of $500,000$ clients of a cell phone company over a three month period. The dataset contains also information about the users' basic demographic data (age, home district, gender, and default status at the end of the three month period) as well as usage information (frequency and duration of calls, messages, and movement records based on frequently used cell towers). We perform traditional network analysis on the data to create new complex features. We analyze weighted links based on the number of communications sent between individuals to reveal the paths between individuals, reciprocity imbalances among users, and community structure in the network.

The default status (an indicator of whether the client stopped paying the phone bill over the course of data collection) is the predicted variable for this study. It is typically accessible as a part of the user's phone records, yet it still measures a complex user behavior combining many different behavioral factors.

To encompass the wide variety of possible correlations between individual attributes, behavioral indicators as well as network metrics and the probability of default, we create an extremely broad feature set, aggregating thousands of features of varying complexity from all the facets of information contained in the CDR dataset. Then, we perform feature contribution analysis and choose the features with the strongest predictive power to balance model performance and its complexity. As discussed in the next section, this general methodology for building complex behavior models is a novel approach and our contribution to the state of the art.

2. Related work on modeling complex human behavior

2.1. Mobile phone data

Call detail records (CDR) is a standard dataset collected by telecommunication operators. For each user, it contains information about telecommunication events in which this user was involved. In recent years, it has become a popular source of information for users' behavioral analytics.

The closest to our goals is the work on modeling credit defaults [22] by building a model of the user's financial risk which yields a score that can be interpreted as the probability of default. This model outperforms the Credit Bureau scores by using thousands of weak predictors derived from CDR and demographic data. This work uses only $60,000$ users and uses only basic level network features such as degree.

2.2. Credit risk management

Our paper focuses on credit risk, which refers to the clients who may stop paying back their loans, e.g., mortgage loan, credit card spending or, in our

case, cell phone bill. Such events are called defaults. Banks and companies traditionally tackle this problem by calculating credit scores or probability of default for each of the potential clients. For individual customers data sets used to predict a customer's probability of default include demographic data, loan and credit information [3], social media [7, 33] or mobile phone data [4].

Traditionally, the logistic model was often used to predict defaults and today is still useful for benchmarking thanks to its simplicity, interpretability and dependability [29, 3, 10]. However, more sophisticated and innovative approaches are also used, like neural networks [29, 33, 11], smart ubiquitous data mining [3], theory of three-way decisions [16], and theory of survival [10]. In our paper, we introduce a novel approach that starts with creating a large number of features (over 6,000 here) and then reducing them to a few well performing subsets.

3. Network Creation and Analysis

In this section, we define the node-level location and communication activity features that indicate how embedded the node is in the network which is likely to determine the social cost of leaving the network. Such social network analysis is common in working with CDR datasets [5], but the highly detailed information contained in the CDR comes with a large amount of noise. Quantifying what level of communication between individuals indicates a connection is challenging. This issue is made worse considering the potential bias introduced by specific patterns of communication, as generational and cultural divides are prominent in phone usage [15]. Therefore use of case-specific methods is common. Some attempts at a more general solution to this problem include reciprocity or activity requirements for links, but these solutions suffer from losing many fine details of the system [12]. More complete results can be obtained by using statistical methods to detect and remove links that are more likely to be random, but the methods come with increased computation cost [14].

In this study for detailed analysis, we use directed graphs in order to preserve the imbalances that tend to arise even among reciprocal relationships [9]. We

	$w_{i,j}$	$w_{i,j} + w_{j,i}$	$\sum_j w_{i,j}$	$\sum_j w_{j,i}$
σ	2.01	2.37	1.47	1.47
β	0.127	0.0959	0.332	0.333
R^2 log-normal	0.9967	0.9991	0.9926	0.9924
R^2 stretched exp	0.9978	0.9994	0.9984	0.9980

Table 1: Fitting parameters and R-squared values for Figure 1. σ corresponds to the fitting parameter in equation 1. β corresponds to the fitting parameter in equation 2.

also use communication frequency between individuals to define edges. We primarily use a weighted network for network analysis and feature generation as it better represents relationship strength and network location properties, but the unweighted network is very useful for understanding many of other interesting properties of the network unrelated to predicting user defaults. For more information on the construction, behavior, and dynamics of unweighted graphs of this network see Supplementary Material Sec. 1.

3.1. Weighted network based on event frequency

Let $w_{i,j}$ denote the number of communications sent from user i to user j is w_{ij}. We choose here the most dense representation of the network in which as long as $w_{i,j} > 0$, a directed link is formed from node i to node j with weight $w_{i,j}$ (some interesting properties of this type of network is discussed in [32, 28, 17]). Figure 1 shows various probability distributions associated with w. These distributions possess a curvature somewhere between an exponential distribution and a power law distribution. It is then appropriate to fit them using a log-normal distribution and a stretched exponential which have the following form.

$$P(w)_{\text{log-normal}} \propto \frac{1}{w} e^{-\frac{(\ln w - \mu)^2}{2\sigma^2}} \tag{1}$$

$$P(w)_{\text{stretched-exp}} \propto e^{-w^\beta/\alpha} \tag{2}$$

Table 1 shows the fitting parameters and R-squared values for Figure 1. Both fits yield high R-squared values.

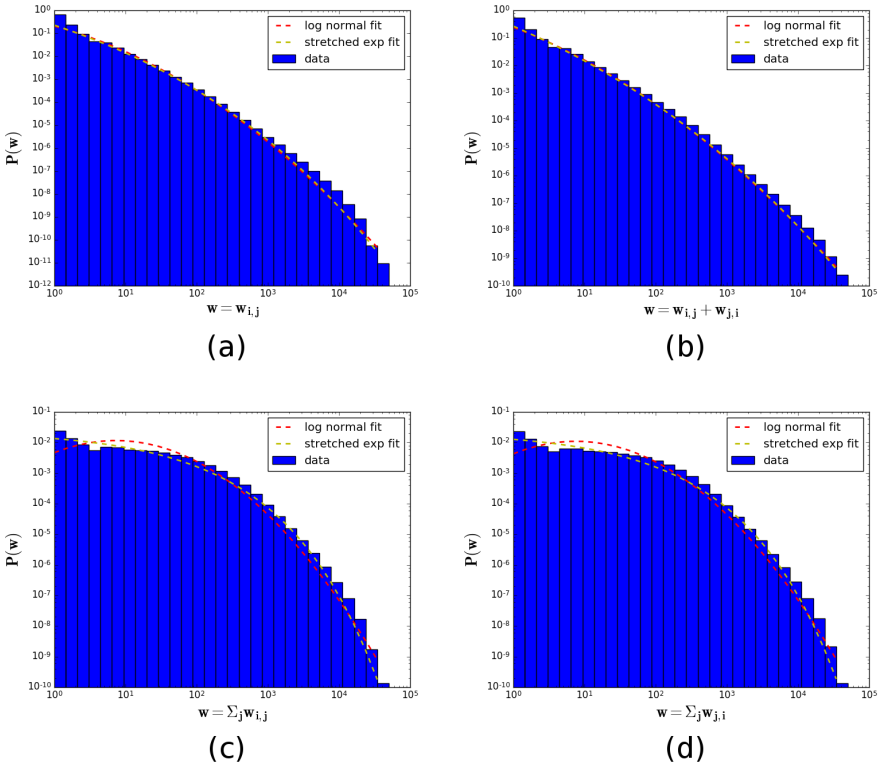

Figure 1: a) Distribution of directional link weights $w_{i,j}$. (b) Distribution of non-directional link weight $w_{i,j} + w_{j,i}$. (c) Distribution of the sum of outgoing link weights of users in the system $\sum_j w_{i,j}$. (d) Distribution of the sum of incoming link weights of users in the system $\sum_j w_{j,i}$. Table 1 shows relevant fit parameters and R-squared values. Both functions in this case reasonably fit the distributions with neither significantly outperforming the other.

We use centrality measures, reciprocity measures, and community detection to define features indicative of how embedded a user is in the network. For all distance-based applications, we use a link's weight to define a normalized distance from the source node i to the target node j, defined as $d_{i,j} = w_{avg}/w_{i,j}$ where w_{avg} is the average weight of all connections in the network. To provide a baseline for comparison, we rewire the network by swapping the edge destinations to create a pseudo-random weighted graphs that maintain the in and out-degree structure of the original network.

We first look at the measure of harmonic closeness centrality (closeness centrality adapted to non-connected graphs) [21, 18]. We define this centrality as $C_H(i) = \frac{1}{N-1}\sum_{j\neq i}\frac{1}{l_{i,j}}$, where N is the total number of nodes and $l_{i,j}$ is the distance of the shortest path between nodes i and j. For the original network, these centrality scores are generally fairly high and evenly distributed, with an average harmonic centrality of $C_{avg}^{cell} = 4.61$ and a standard deviation of $C_{std}^{cell} = 1.84$. Both values are higher than those for the randomly rewired graph which yields $C_{avg}^{rand} = 4.11$ and $C_{std}^{rand} = 1.20$. Interestingly, the diameter and average shortest path length of the giant component in the original network $D^{cell} = 6.24$ and $\langle l^{cell}\rangle = .311$ are also slightly higher than the corresponding values $D^{rand} = 4.35$ and $\langle l^{rand}\rangle = .30$ for the randomly rewired network.

These differences reveal the basic shape of the original network, which is characterized by significant populations of both highly connected and highly isolated nodes. This property is shown clearly via the diameters of the two networks. The original network is significantly wider than its randomly rewired counterpart. Despite the existence of close communities and hubs within the original network, there are multiple extremely remote nodes in it with no close ties.

We also use this weighted network to measure reciprocity [28], which is used to measure how one sided interactions are. Our first measure is how many pairwise communications are matched with communications in the opposite direction. A surprisingly low number of communications, 62.88%, have communications in both directions indicating that over a third of interactions are unreciprocated.

Next, we construct a reciprocity metric that measures the average contribution of a nodes' links so that it is independent of the degree of the node. Unlike [28], we apply it to directional links as opposed to [28]. The corresponding definition is $R_i = \frac{1}{k_i}\sum_{j\in N(i)}\frac{w_{i,j}-w_{j,i}}{w_{i,j}+w_{j,i}}$, where k_i is the total (both in and out) degree of node i, $w_{i,j}$ is the number of communications from node i to node j, and $N(i)$ is the set of neighbors that have a link connected to node i. According to this definition, links over which a node is sending more than receiving

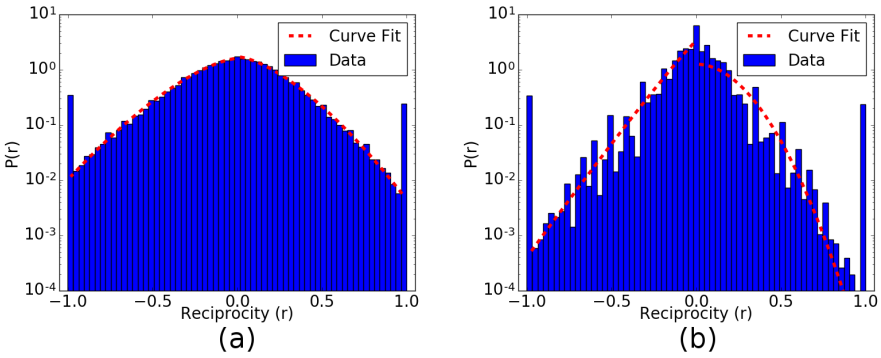

Figure 2: a) The distribution of reciprocities. (b) The distribution of reciprocities using a randomized network. Peaks are observed at -1, and 1. This is consistent with a large number of nodes have low degree as most links are not reciprocated. The randomized network produces a narrower and rougher distribution indicating some level of order in how individuals interact.

contribute positively to the node's metric, while links with the reverse pattern contribute negatively. The absolute value of the contribution itself increases monotonically with the difference in the level of communication [28]. Finally, every link's contributions are normalized, thus bound within the range [-1,1].

We show in Fig 2(a) that the distribution of reciprocity is fairly smooth with large spikes at the extreme values. These spikes arise because there are nodes with either no incoming communications ($R_i = 1$) or no outgoing communications ($R_i = -1$). Since all data comes from a single cell phone provider, it is likely that many of these nodes simply have contacts that use different carriers not included in the set. Removing these outliers, we fit the remaining distribution using a stretched exponential as described in Equation 2. Table 2 shows the fitting parameters for Fig 2. As can be seen, there is an asymmetry in the width of the positive and negative side of the distribution. This can be caused by an increased amount of communication for nodes with reciprocity equal to 1. While each link's contribution $w_{i,j} - w_{j,i}$ is symmetric, the overall contributions are averaged and thus the symmetry can be broken. If an unreciprocated link is the only link for a node, its reciprocity will be $R_i = \pm 1$ regardless of what w is. This means if there are overall more communications from nodes with

Fitting Parameter	CDR Graph		Random Graph	
	Positive	Negative	Positive	Negative
β	1.52	1.49	1.96	1.07
α	0.168	0.198	0.080	0.111
R^2	0.994	0.991	0.687	.961

Table 2: The fitting parameters used in Fig 2. Positive and Negative refer to the fit of the distribution where the reciprocity values are positive and negative respectively. Due to the non-integer value of the curvature, we fit the absolute value of the negative side and plot accordingly. For the CDR graph both positive and negative sides have very similar fitting parameters except for the width in the distribution which suggest an asymmetry.

$R_i = 1$ than nodes with $R_i = -1$, this difference would show itself in the overall distribution.

For the randomized network, Fig 2(b), the concavity of the distribution narrows compared to Fig 2(a) from the presence of more frequent values close to zero. Hence, the original network is more diverse than its randomized counterpart is.

For other possible representations of reciprocity within the network, see Supplementary Material Sec. 2, where we discuss two additional variants. These alternative metrics are not independent from the metric presented in this section.

3.2. Community detection and geographical districts

The CDR based network allows us to analyze the social communities present in the system. We use the GANXiS(SLPA) algorithm for its ability to detect even disjointed and overlapping communities and fully encapsulate the social structure of the network [31, 30]. Using this algorithm, we identify a set of over 6450 social communities, many more than the 231 geographic communities derived from the districts reported. Despite the size differences (the largest district contains 63491 users while the largest social community has just 741 with an average of only 74.65 users), the groups are substantial enough to test the overlap of the lists for greater insight into how the social ties form. Intuitively, it

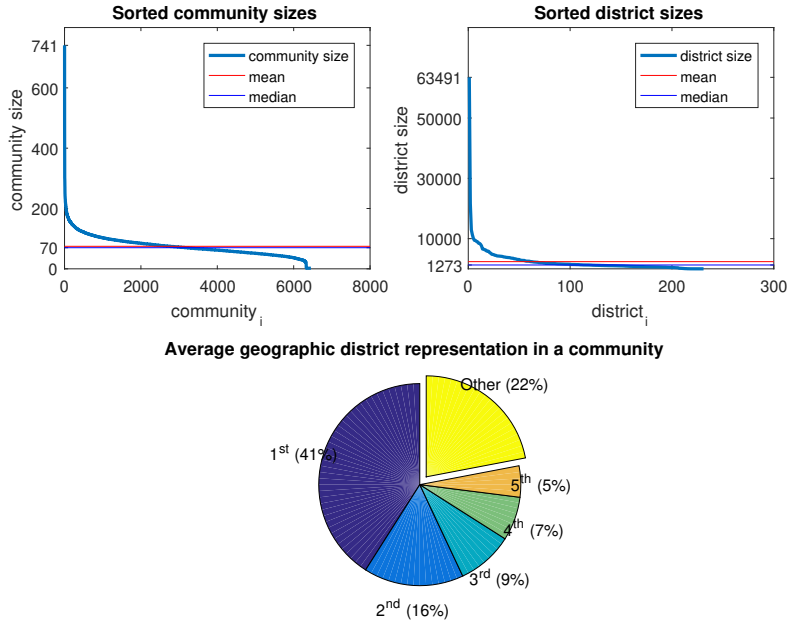

Figure 3: Top: Community and districts sizes, which are sorted in descending order. The average social community size is 75 with a median of 70, whereas geographic districts show an average of 2,380 and median of 1273. Bottom: average proportion of users from a community that belong to the same geographic district - from most represented geographic district by users (1^{st}) to the 5^{th} most represented.

seems reasonable to expect that the social communities are highly influenced by the geographic district of their members. Instead we see in Fig. 3 that on average only 41% of each community comes from the same district, and in fact even the top five districts only account for 78% of each community's makeup. The diversity of the geographic locations within social groups is especially surprising given the generally small size of the social groups compared to the geographic districts. The communities are also included in our feature set for predicting user default's as it is possible that individuals of a certain groups are more likely to default.

4. Feature Generation

We combine the above described network features with the various raw usage and location features that are inherent to the dataset to create our full feature set. For ease of analysis, we divide them into six groups:

1. **High Level Network features -** explained in detail above. This subset consists of 5 features.
2. **Consumption features -** includes information about the total number of communication events, the total and average duration of phone calls, and average time between consecutive communications. This subset consists of 2784 features.
3. **Correspondent features -** based on the distinct number of individuals that each user communicates with over various time periods. This subset consists of 2561 features.
4. **Reciprocated event features -** number of events where the observed user returns a call/message within a specified time period of receiving a communication. This subset consists of 672 features.
5. **Mobility features -** includes the movement patterns of individuals based on the cell tower used for each communication with relation to commonly used towers. This subset consists of 29 features.
6. **Location features -** includes the two most used cell phone towers for each user. This subset consists of 2 features.

In some cases large or very inclusive features can be broken up by analyzing time windows (hours, days, day of the week, weeks, months, business hours, non-business hours, weekend, weekday), direction (ingoing/outgoing event), and communication type (call/message), leading to a large overall set of more than 6000 features. Most of these features are standard, so we do not present them in detail here, but relegate the full descriptions of each feature set to Supplementary Material Sec. 3.

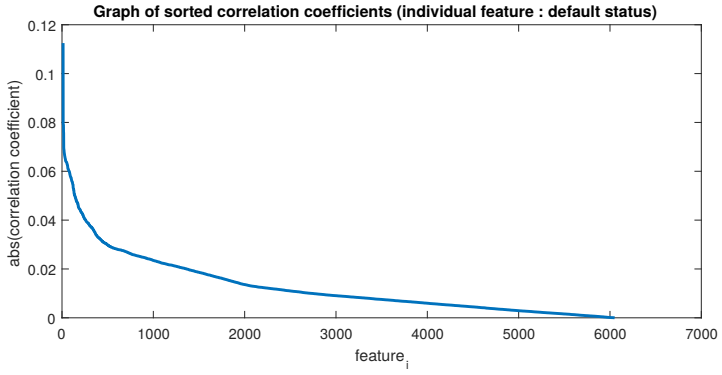

Figure 4: This graph shows point biserial correlation of each individual feature to default status. Values are sorted in descending order by absolute value of their correlation coefficient. Features with highest absolute correlation (originally all negative) are the average number of used cell towers in one week, distance traveled in a day of week and the average daily radius.

5. Predicting default status

5.1. Correlation analysis

We utilize point biserial correlation to define the basic correlation between each feature and user default status due to its suitability for handling both categorical and continuous variables as are present in our feature set. This process indicates that individual features tend to have very little correlation to default status (shown in Fig. 4), with a maximum absolute correlation of only 0.1155 and average correlation of -0.006. This poor individual correlation shows that no single feature reproduces the default status of the individuals reliably, proving the value of building an accurate predictive model through more complex analysis.

5.2. Modeling

5.2.1. Logistic regression

To create the predictive models, we start by defining the notation for logistic regression. Let β_0 denote the intercept and β stand for a vector of regression coefficients of length p. Both are determined by maximum likelihood estimation method which uses a training set with n known outcomes in the vector Y for

nodes in the training set. Matrix X of size $n \times p$ with rows corresponding to nodes and columns corresponding to features contains values of features for all nodes in training set. For any node u in test set and its corresponding vector of feature values x_u of size p, the probability of "success" (in our case the default status of this node being 1) is defined as

$$P(y_u = 1|x_u) = \frac{1}{1 + \exp^{-\beta_0 - \beta \cdot x_u}}, \tag{3}$$

where y_n denotes the value of the dependent variable for node u.

5.2.2. Principal Component Analysis

To build a more complex model that utilizes the other 6048 features, we first normalize the features to mean values of 0 and variance 1. Once normalized, to deal with the extremely large number of features we perform Principal Component Analysis [19](PCA) to decompose the feature space and yield a set of "principal components". The results are linearly uncorrelated and ranked in such a way that the first principal component explains the highest possible amount of variability in the data, while each following component explains the highest amount of variance under the condition that it is orthogonal to all preceding components. Once these components are obtained, we select subsets of the components that account for large amounts of variability in the data and use them as explanatory variables for a linear model. We do this for both large and small subsets of the components, as pruning the data in this way represents a careful balance between model simplicity and accuracy. In our dataset, this balance manifests in the rapidly diminishing returns seen from larger component inclusion. The first component explains 20% of variability in the data while the second explains only 7%. In fact, the first thirty components together explain only 42% of the variability, while the first five hundred sum up to just 66%. For this reason, we generate two distinct logistic models: a simple one based on the first 30 PCA components (*pca-30*) and a more complex one, based on the first 500 PCA components (*pca-500*). We use notation $p1$ for the number of selected principal components. We then map our features into the new space to fit these

groupings, such that

$$X_{PC} = X \cdot C,$$

where $C \in \mathbb{R}^{p \times p_1}$ is the matrix of p_1 principal components and $X_{PC} \in \mathbb{R}^{n \times p_1}$ is the new feature matrix. Finally, we also fit a model (*pval-05*) using only those variables from *pca-500* that have a $p - value < 0.5$ to include the greater detail of the larger model with a lower complexity for calculation.

5.2.3. Other models

We further examine the data by utilizing other techniques, and build models for each as a comparison to the above described PCA models. For instance, one large issue with our dataset is that positive samples (defaulted users) account for only 0.25% of the whole dataset, making it extremely unbalanced. Learning from this kind of unbalanced dataset is a well-known challenge in the data mining community that we attempt to account for by performing oversampling (multiplying positive samples to make them a larger portion of the dataset) [6]. In this case, we use multiplication factors ranging from $2-100$, but in the model benchmarks presented here, we show only the model with a multiplication factor of 2 (*oversampled-2*) since it performed the best. Oversampling was applied to the reduced dataset (X_{PC}).

Further, as an alternative to the PCA reduction presented in the prior section, we also build a separate model by selecting features via Lasso regression [26]. This method adds a penalty term to the log likelihood function in the prediction that shrinks the coefficients of less important variables to zero. We fine tune this reduction via a free parameter λ, varying it through a range of values to obtain the best lasso fits. The efficacy of this method is demonstrated in two separate models: one based on logistic regression (*lasso-logistic*) and another using a Support Vector Machine [27] (*lasso-svm*).

Finally, the last model we use for comparison utilizes a simple method for feature space reduction that aggregates some of the more specific features into general descriptions of behavior. For instance, features that were based on a particular day or week of the three month period are grouped and combined,

creating instead features for weekdays versus weekend averages or business hours versus off-hours. This aggregation not only simplifies the model, but also makes it generalizable to other datasets with different levels of detail. This method shrinks the overall feature set considerably, leaving only 781 features. As before, these features are then normalized before being further reduced via PCA.

6. Results and Discussion

6.1. Experimental setting

After each of the above models is fit to a training set comprising 70% of the whole dataset, they are tested on the remaining 30% of the data and evaluated via their recall, fall-out, and precision. Defaulting users are labeled as positive examples and default predictions are defined as those users within the 95^{th} percentile of default probabilities. This threshold is chosen to be relatively low to fit the nature of our study where false positives (non-defaulted users that were predicted as defaulters) are less damaging than false negatives (unidentified defaulters) to companies.

The recall (the true positive rate) is the rate of identified defaulting customers. The fall-out (the false positive rate) is the probability of labeling a good client as a defaulting one. The precision is the fraction of correct default predictions out of all default predictions. These metrics are calculated as

$$recall = \frac{TP}{TP + FN}, \quad fall\text{-}out = \frac{FP}{FP + TN}, \quad precision = \frac{TP}{TP + FP}$$

where TP is the number of true positives, FN is the number of false negatives, FP is the number of false positive, and TN is the number of true negatives.

6.2. Evaluation

The performance of the models tested can be compared using a receiver operating characteristic curve (ROC), which we show in Figure 5. Here, we focus this comparison on the recall and fall-out of the model, where recall (which identifies defaulting customers) is of interest to phone companies, while fallout

has greater relevancy in other applications of the models. The exact values of these performance metrics can be seen in Table 3, which also shows the precision of each model for reference.

The initial results are in line with what would be expected intuitively: the worst performance comes from the random model, followed by *glm-7* (the logistic model based on only 7 features). As the models become more complicated, their performance tends to increase. For instance, a significant reduction in false positives can be seen in the *lasso-logistic model* (which uses 309 variables) over the *glm-7*, then a further reduction in the *lasso-svm* model that uses 475 variables. Additionally in both Lasso models, the variable with the highest coefficient is the user's most commonly utilized cell tower, indicating the presence of geographic regions which are high risk areas for user defaults. Other high performance models include the various PCA models, led by *pca-500*, *pval-05* and *oversampled-2* with *pval-05* outperforming all other models.

From these results, it is clear that for the most part larger feature sets allow for more accurate models (as would be expected), but this is not a strict rule. The best performing model is the one that begins with a large sample of features, but is then stripped of those that aren't considered significant. In other words, there are likely some 'false flags' in the feature list that tend to confuse the model rather than contribute. Feature removal only works to a point, however, as the more aggressive methods in the *pca-aggr* model lose these benefits and in fact make it one of the worst performing models tested. Thus, there is likely some very meaningful information even within the extremely specific features that would be intuitively too limited to contribute much. Finally, it should be noted that while the highest performing oversampling model, *oversampled-2* yields improvements comparable to filtering insignificant features, applying both techniques *worsens* the results as the model apparently over-fits.

6.3. Stability

For a more in depth look into the differences in prediction among the top three models (*oversampled-2*, *pca-500*, *pval-05*), we look at the overlap size

Model	Recall	Fall-out	Precision
random	0.060	0.0501	0.003
glm-7	0.224	0.0495	0.012
lasso-logistic	0.484	0.0490	0.023
pca-aggr	0.676	0.0484	0.036
lasso-svm	0.749	0.0482	0.040
pca-30	0.810	0.0480	0.043
pca-500	0.889	0.0478	0.047
oversampled-2	0.897	0.0478	0.047
pval-05	**0.900**	**0.0477**	**0.048**

Table 3: Recall, fall-out and precision for each of the models presented in Figure 5. Precision is low due to the fact that the dataset is majorly unbalanced; however, precision of the best model is about 15 times higher than in the random model.

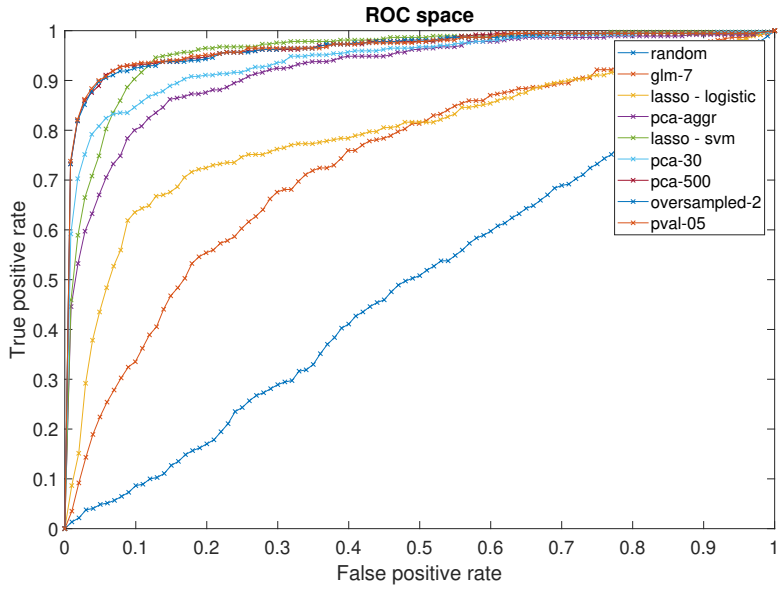

Figure 5: Model performance when predicting defaulting users from worst to best: *glm-7*, *lasso-logistic*, *pca-aggr*, *lasso-svm*, *pca-30*, *oversampled-2*, *pca-500* and *pval-500*.

(the intersection of correctly labeled defaulting customers) and average overlap score (the similarity of two rankings at increasing depth giving higher weights to higher ranked observations; AOS). All analysis is restricted to only those customers with a calculated probability of default (PD) in the $95-th$ quantile. As can be seen in Table 4, both the overlap and AOS are extremely high for all model comparisons, indicating that all three models are stable and predict roughly the same nodes for default. Further, the *pval-05* and *pca-500* models can be seen to be more similar to each other than the *oversampled-2* model due to *pval-05* being essentially *pca-500* with some of the problematic features removed.

		pval-05	pca-500	oversampled-2
pval-05	overlap	100	98.8	97.3
	AOS	100	96.5	94.3
pca-500	overlap		100	97.9
	AOS		100	95.5
oversampled-2	overlap			100
	AOS			100

Table 4: Overlap size (in percentage) and average overlap score of correctly labeled defaulting customers for the three best models.

6.4. Contribution of feature sets

Finally, we identify the original features that contribute the most to our models to gain a better understanding of what aspects of the mobile phone data are most important for predicting default status. However, as we mentioned before in Sec. 5.1, each feature is generally fairly poorly correlated with defaults. Thus it is no surprise that there is no one outstanding feature that stands out as the strongest contributor. To remedy this, we perform a general analysis using the feature set groupings described in Sec. 4 to identify which class of features contributes the most. By building a host of modified models, each consisting of all of the feature sets except one, we can compare the performance of the

Model	Recall	Precision	Δ_{Recall}	$\Delta_{Precision}$
Full feature set	0.9	0.048	-	-
- consumption features	0.88	0.047	0.02	0.001
- correspondent features	0.58	0.031	**0.32**	**0.017**
- reciprocated features	0.88	0.046	0.02	0.002
- mobility features	0.88	0.046	0.02	0.002
- network features	0.88	0.05	0.02	-0.002
- cell tower PD features	0.89	0.050	0.01	-0.002
Only correspondent features	0.86	0.049	-	-

Table 5: Results of models based on reduced feature sets. In each row, the underlying dataset is missing one category of features (listed in first column of the table). Two models that proved to be the best were fitted in each case: *glm-500* and *glm-500-filt*. Most features contribute only about 2-3% to the final recall and 0.1-0.2% to the final precision, except for correspondent features which contribute 33% of recall and 1.9% of precision.

models with the reduced data against the model with all data included. This method allows us to gain a better understanding of how much information is lost when a feature set is removed (i.e. how unique that information is), data that is especially valuable to this study considering the high redundancy in our data set.

As shown in Table 5, by far the highest contributors to recall and precision are the correspondent features, which focus on individual's unique frequent correspondents in various time frames. To better understand exactly what features within the correspondent feature set are important, we perform 5-fold cross validation on a linear model using only these features. On average, this model is able to achieve a 0.86 recall and 0.049 precision, both fairly high for such a small subset of all features. Further, we explore which out of the 2543 correspondent features contribute most to the final PD by mapping the maximum likelihood estimation (MLE) coefficients of PCA components back to original features' coefficients,

$$\hat{\beta} = C \cdot \hat{\beta_{PC}}$$

where $\hat{\beta} \in \mathbb{R}^{2543\times1}$, $C \in \mathbb{R}^{2543\times500}$ is the matrix of principal components and $\hat{\beta}_{PC} \in \mathbb{R}^{500\times1}$ are the MLE coefficients for a model fitted on 500 PCA components.

We then multiply each regression coefficient by the mean of the corresponding feature (separately for paying and defaulting customers). Finally, we define a score for each user based on the values of $\hat{\beta}X_j$, $j = 1, ...n$. This creates a system where higher scores correspond to customers with higher PD's, and leads to an average score of -0.0453 for paying customers and 16.2758 for defaulting customers. This analysis is not possible using the whole feature set together, as the normalization of the features leads to an average contribution of zero, but the absolute contribution values of each feature are taken into account to maintain a focus on the actual impact of the features on the PD. The results for the four strongest features are presented in Table 6. In the end, the same ten features are identified as the strongest contributors for both the paying and defaulting customer groups. All ten relate to the number of unique correspondents with whom the user communicates during the holiday period around Christmas and New Year's Eve. This strongly indicates a tie between not only unique correspondents and defaulting, but also unique correspondents around holiday periods (when users are most likely to be contacting close family and friends). From these results we can begin to draw a clear connection of low probability of default with the large number of active unique ties an individual has within the network, and the high strength of those ties.

However, it should be noted that this kind of estimation of variable contribution is too simple to provide a definitive insight into the relative importance of the variables. There are several other approaches that compare predictors in regression including methods considering variable importance via R^2 partitions [20, 24], dominance analysis [2] and relative weights analysis [8, 13]. According to the literature, the dominance analysis provides the most accurate results, but it is not practical for our problem since it measures relative importance in a pairwise fashion and is suitable only for small variable sets. Instead, the relative weights approach is sometimes used for systems with multiple

Feature	Mean relative contribution
unique message correspondents (12/24)	0.043
unique message correspondents (incoming, 12/24)	0.041
unique message correspondents (outgoing, 12/24)	0.024
unique correspondents (outgoing, 12/24)	0.023
sum	0.131

Table 6: Four features with the highest contribution to the PD calculated on separate sets of paying and defaulting customers. All of the top 10 relate to the number of unique correspondents during holiday period around Christmas and New Year's Eve.

correlated predictors as can be found here [13], but this approach is theoretically flawed and is therefore not recommended for use [25]. We instead utilize a variable importance (VI) extension for logistic regression [24] that is based on Pratt's axiomatic [20] and the geometric approach of Thomas [23]. This method equates VI to variance explained by the variable, which is $\beta_j \rho_j$, where β_j is the standardized regression coefficient and ρ_j is the simple correlation between variables Y and X_j. For our purposes, we use the geometric approach to this method; an interpretation of Pratt's measure based on the geometry of least squares. Here, the VI indices are defined as

$$d_j = \frac{\hat{\beta}_j \hat{\rho}_j}{R^2}, j = 1, \ldots, p, \tag{4}$$

where hats denote sample estimates, and R^2 is as usual the proportion of sample variance explained.

Further, we utilize a pseudo-R^2 measure based on Weighted Least Squares such that this set of indices sums to one and the importance of a subset of variables is equal to the sum of their individual importance. This results in the four most important features, according to VI metric, shown in Table 7, yielding results very similar to the ones presented in Table 6. The two most important variables are: the number of unique correspondents with whom the user had a call on Dec 24th with relative importance of 0.048, and the number of unique correspondents from which the user received a message on Dec 24th

Feature	d_j
unique message correspondents (12/24)	0.048
unique message correspondents (incoming, 12/24)	0.045
unique message correspondents (outgoing, 12/24)	0.029
unique correspondents (outgoing, 12/24)	0.026
sum	0.148

Table 7: Ten features with the highest contribution to the PD. All of them relate to the number of unique correspondents during holiday period around Christmas and New Year's Eve.

with relative importance of 0.045. This finding reiterates the significance of the number of unique correspondents with whom the user communicates around holiday periods. Finally, to ensure that these features actually carry unique information, we fit another model removing from the dataset the features with the highest identified VI. Doing so reduces the recall by 0.042 and leaves none of the new identified important variables with a relative importance above 0.03.

7. Conclusion

In this paper, we investigate many different aspects of user behavior to build a large suite of features for analysis, starting with constructing the underlying social networks based on the cell phone usage data. This produces a complex network rich with information. Applying common network metrics reveals some of the characteristics of the network such as the heterogeneity of the centrality, diameter, and reciprocity measures.

The complex nature of the problem addressed by our machine learning method to predict whether users will default in paying their cellphone bill leads us to utilize some 6000 features, which we then pare down to only the most predictive ones. The resulting model achieves a recall of 0.9 with a fall-out of only 0.048, the performance that compares favorably with [22, 1], e.g., recall of 0.674 is reported in the later reference.

Finally by investigating in depth the various features that contribute to the model, we are able to pinpoint the surprisingly best, contributor: the number of unique contacts with whom the user interacted around the winter holidays (when users are most likely to contact their closest friends and family). The significance of this correspondent information is higher than that of more traditionally used features. These results demonstrate the need for systematic approach to selecting features for complex behavior prediction. Indeed, our results show that the strength of links within the network is better determined by the timing of communications rather than the volume, duration, or distance traditionally used for the similar predictions.

While our use of CDR data here focuses on user analytics and predictive modeling, CDR datasets can support many other avenues of research. This work investigates the default status of individuals, but there are many other complex aspects of user behavior that could benefit from similar computational techniques. The modeling we present could be improved by gathering information over larger time periods in order to get a larger population of defaulting individuals. Finally, with new machine learning algorithms constantly being designed and improved, a more specific algorithm that is built for high dimensional data such as ours could improve the resultant predictions and understanding of how the specific features contribute to the overall model which is novel and unexpected. Hence, our novel approach leading to these results is the main contribution of our paper to the state of the art.

8. Acknowledgments

Funding: This work was supported in part by the Army Research Laboratory (ARL) under Cooperative Agreement Number W911NF-09-2-0053 (NS-CTA), by the Office of Naval Research (ONR) Grant No. N00014-15-1-2640, and by RENOIR EU H2020 project under the Marie Skłodowska-Curie Grant Agreement No. 691152. 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.

References

[1] Agarwal, R. R., Lin, C.-C., Chen, K.-T., and Singh, V. K. (2018). Predicting financial trouble using call data—on social capital, phone logs, and financial trouble. *PloS one*, 13(3):e0191863.

[2] Azen, R. and Budescu, D. V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological methods*, 8(2):129.

[3] Bae, J. K. and Kim, J. (2015). A personal credit rating prediction model using data mining in smart ubiquitous environments. *International Journal of Distributed Sensor Networks*, 11(9):179060.

[4] Björkegren, D. and Grissen, D. (2017). Behavior revealed in mobile phone usage predicts loan repayment. *arXiv preprint arXiv:1712.05840*.

[5] Blondel, V. D., Decuyper, A., and Krings, G. (2015). A survey of results on mobile phone datasets analysis. *EPJ Data Science*, 4(1):10.

[6] Chawla, N. V. (2009). Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook*, pages 875–886. Springer.

[7] Ge, R., Feng, J., Gu, B., and Zhang, P. (2017). Predicting and deterring default with social media information in peer-to-peer lending. *Journal of Management Information Systems*, 34(2):401–424.

[8] Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate behavioral research*, 35(1):1–19.

[9] Kovanen, L., Saramaki, J., and Kaski, K. (2011). Reciprocity of mobile phone calls. *Dynamics of Socio-Economic Systems*, 2(2):138–151.

[10] Kuznetsova, N. V. and Bidyuk, P. I. (2017). Modeling of credit risks on the basis of the theory of survival. *Journal of Automation and Information Sciences*, 49(11).

[11] Kvamme, H., Sellereite, N., Aas, K., and Sjursen, S. (2018). Predicting mortgage default using convolutional neural networks. *Expert Systems with Applications*, 102:207–217.

[12] Lambiotte, R., Blondel, V. D., de Kerchove, C., Huens, E., Prieur, C., Smoreda, Z., and Van Dooren, P. (2008). Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317–5325.

[13] LeBreton, J. M. and Tonidandel, S. (2008). Multivariate relative importance: Extending relative weight analysis to multivariate criterion spaces. *Journal of Applied Psychology*, 93(2):329.

[14] Li, M.-X., Palchykov, V., Jiang, Z.-Q., Kaski, K., Kertész, J., Miccichè, S., Tumminello, M., Zhou, W.-X., and N Mantegna, R. (2014). Statistically validated mobile communication networks: the evolution of motifs in European and Chinese data. *New Journal of Physics*, 16(8):083038.

[15] Ling, R., Bertel, T. F., and Sundsøy, P. R. (2012). The socio-demographics of texting: An analysis of traffic data. *New Media & Society*, 14(2):281–298.

[16] Maldonado, S., Peters, G., and Weber, R. (2018). Credit scoring using three-way decisions with probabilistic rough sets. *Information Sciences*.

[17] Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., and Barabási, A.-L. (2007). Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences of the United States of America*, 104(18):7332–6.

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

[19] Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

[20] Pratt, J. (1987). Dividing the indivisible using simple symmetry to partition variance explained. *Proceedings of the Second International Conference in Statistics*, pages 245–260.

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

[22] San Pedro, J., Proserpio, D., and Oliver, N. (2015). Mobiscore: towards universal credit scoring from mobile phone data. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 195–207. Springer.

[23] Thomas, D. R., Hughes, E., and Zumbo, B. D. (1998). On variable importance in linear regression. *Social Indicators Research*, 45(1-3):253–275.

[24] Thomas, D. R., Zhu, P., Zumbo, B. D., and Dutta, S. (2008). On measuring the relative importance of explanatory variables in a logistic regression. *Journal of Modern Applied Statistical Methods*, 7(1):4.

[25] Thomas, D. R., Zumbo, B. D., Kwan, E., and Schweitzer, L. (2014). On johnson's (2000) relative weights method for assessing variable importance: A reanalysis. *Multivariate behavioral research*, 49(4):329–338.

[26] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

[27] Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999.

[28] Wang, C., Lizardo, O., Hachen, D., Strathman, A., Toroczkai, Z., and Chawla, N. V. (2013). A dyadic reciprocity index for repeated interaction networks. *Network Science*, 1(1):31–48.

[29] Wang, Y. and Lee, Y.-C. (2018). Customer credit evaluation using big data of microfinance company in china. *The Korean Academic Society of Business Administration*, (2):1601–1645.

[30] Xie, J. and Szymanski, B. K. (2011). Community detection using a neighborhood strength driven label propagation algorithm. In *IEEE NSW 2011*, pages 188–195.

[31] Xie, J., Szymanski, B. K., and Liu, X. (2011). Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. In *ICDM 2011 Workshop on DMCCI*.

[32] Yan, X., Jeub, L. G., Flammini, A., Radicchi, F., and Fortunato, S. (2018). Weight thresholding on complex networks. *Phys. Rev. E 98, 042304*.

[33] Zhang, X.-j. and Hu, J. (2009). Personal credit rating assessment for the national student loans based on artificial neural network. In *Business Intelligence and Financial Engineering, 2009. BIFE'09. International Conference on*, pages 53–56. IEEE.