Journal of the Operational Research Society (2016) 67, 1135–1145 © 2016 Operational Research Society Ltd. All rights reserved. 0160-5682/16

www.palgrave.com/journals

Predicting time-to-churn of prepaid mobile telephone customers using social network analysis

Aimée Backiel[1]*, Bart Baesens[1,2,3] and Gerda Claeskens[1]

[1]Katholieke Universiteit Leuven, Leuven, Belgium; [2]University of Southampton, Highfield Southampton, UK; and [3]Vlerick, Leuven-Gent Management School, Gent, Belgium

Mobile phone carriers in a saturated market must focus on customer retention to maintain profitability. This study investigates the incorporation of social network information into churn prediction models to improve accuracy, timeliness, and profitability. Traditional models are built using customer attributes, however these data are often incomplete for prepaid customers. Alternatively, call record graphs that are current and complete for all customers can be analysed. A procedure was developed to build the call graph and extract relevant features from it to be used in classification models. The scalability and applicability of this technique are demonstrated on a telecommunications data set containing 1.4 million customers and over 30 million calls each month. The models are evaluated based on ROC plots, lift curves, and expected profitability. The results show how using network features can improve performance over local features while retaining high interpretability and usability.

Journal of the Operational Research Society (2016) **67**(9), 1135–1145. doi: 10.1057/jors.2016.8;
published online 16 March 2016

Keywords: decision support systems; telecommunications; churn prediction; social network analysis; survival analysis

Introduction

The availability of no-contract mobile telephone subscriptions and convenient number portability make it easier for customers to switch and more interesting for companies to accurately predict likely churners. Churn prediction is an established application of data mining in which historical data about previous customers can be used to classify current customers as likely churners or not (Verbeke *et al*, 2012). Traditional churn prediction models rely on local customer attributes, however this can often be incomplete for prepaid subscriptions, which can be purchased and used anonymously. To overcome this, social network analysis is proposed to build models based on calls between customers rather than customer attributes. Customers are not independent, their behaviour depends on the behaviour of those around them (Zhang *et al*, 2012). Using data from a large Belgian mobile carrier, this research demonstrates one application of social network analysis to predict churn and outperform the local prediction model.

With a saturated wireless market, acquiring new customers is difficult and providers must focus more on retaining current customers. Bersen *et al* (2000) estimated that the average churn rate for mobile customers was 2.2% per month. The data set used in this paper indicates a churn rate of 2.9% per month. It has been estimated that attracting a new customer may be five to six times more expensive than retaining a customer (Rosenberg and Czepiel, 1984). In addition, established customers are less sensitive to competition and generate greater profits at lower costs (Verbeke *et al*, 2011). As profitability is the driver behind customer retention, the models in this study will be evaluated using the expected maximum profitability measure of Verbraken *et al* (2013).

Existing literature approaches the problem of retaining customers from two perspectives: explanatory to identify determinants of churn and predictive to identify likely churners. Kim and Yoon (2004) investigated the determinants of churn based on a survey of the Korean mobile market. They found that service attributes (call quality, tariff level, handsets, and brand image) and subscription duration all reduced the probability of churn, while number portability makes churn more likely. Similarly, marketing literature links customer loyalty, customer satisfaction, and switching costs to decreased churn and greater profitability (Lam *et al*, 2009). Strengthening customer loyalty by focusing on service attributes that meet customer demands should be an integral part of the business strategy (Lam *et al*, 2009; Huang *et al*, 2012; Aksoy *et al*, 2013). However, it is not always feasible to improve service attributes. More targeted campaigns may have a better return for the carrier. Such a campaign involves ranking customers according to their probability of churn and sending promotions, such as a bonus credit, directly to those at highest risk in an effort to discourage the potential churn. The bonus has an associated cost, but in exchange the customer will extend their

*Correspondence: Aimée Backiel, KU Leuven—Leuven Institute for Research in Information Systems, Naamsestraat 69—bus 3555, Leuven 3000, Belgium.
E-mail: aimee.backiel@kuleuven.be

Table 1 Churn prediction literature

Churn Prediction Literature	*Data*		*Classifiers*									*Evaluation*		
	Observations	*Features*	*DT*	*RB*	*Reg*	*NB*	*SVM*	*NN*	*Ens*	*Cox*	*Rel*	*AUC*	*Lift*	*TM*
Ultsch (2001)	300000	21												
Wei and Chiu (2002)	114000	6	✓										✓	✓
Dasgupta *et al* (2008)	2100000	2									✓		✓	✓
Pendharkar (2009)	195956	4						✓					✓	✓
Lima *et al* (2009)	5000–10000	20–21	✓	✓	✓							✓		✓
Richter *et al* (2010)	>16000000	3									✓		✓	
Risselada *et al* (2010)	100000	11	✓		✓								✓	✓
de Bock and den Poel (2011)	3827–43305	15–529	✓						✓			✓	✓	✓
Verbeke *et al* (2011)	5000	21		✓	✓			✓						✓
Wong (2011)	4896	7								✓				
Chen *et al* (2012)	633–8842	3–14					✓					✓	✓	✓
Huang *et al* (2012)	827124	30	✓		✓	✓	✓	✓				✓		
Verbeke *et al* (2012)	2180–338924	15–727	✓	✓	✓	✓		✓	✓			✓		✓
Zhang *et al* (2012)	>1000000	39	✓		✓			✓				✓	✓	✓
Abbasimehr *et al* (2013)	10000	171		✓					✓					✓
Verbraken *et al* (2013)	2180–338924	15–727		✓	✓			✓				✓		✓
Verbeke *et al* (2014)	>1000000	2									✓		✓	

Classifiers: DT—Decision Trees, RB—Rule-Based, Reg—Regression, NB—Naive Bayes, SVM—Support Vector Machines, NN—Neural Networks, Ens—Ensemble Methods, Cox—Cox Regression, Rel—Relational Learner
Evaluation: AUC—Area Under the ROC Curve, TM—Threshold Metric (eg classification error, true positive rate, etc)

commitment for some period of time. Deciding which customers should receive such a promotion is the main topic of the remainder of this paper.

Nitzan and Libai (2011) studied the impact of social effects on customer retention and found an 80% increase in churn hazard after the churn of a neighbour. Expanding on this idea, this paper aims to investigate how social social network analysis can be incorporated seamlessly into a churn prediction model and seeks to assess any improvements in accuracy, profitability, and timeliness. Prediction models were built using customer attributes, network attributes, and a combination of both. Models trained using network features were compared with similar models including only local features in order to assess the added value of the derived network features. Three types of models were employed: Cox proportional hazards, logistic regression, and artificial neural networks. Following this introduction, the paper is organized into four remaining sections. In Background, the existing literature on churn prediction in the telecommunications domain will be reviewed. In Methodology, the data and analysis techniques used in the course of this study are described. The results of the experiments are analysed in Findings. Finally, the Conclusion includes a summary of this study, as well as limitations and areas of future research.

Background

Customer churn prediction is widely researched because of its suitability to classification models and its applicability in business. The scope of this literature review is limited to the telecommunications domain, though the churn prediction process will be similar in most settings. After data is collected, it must be labelled with the outcome variable in this case churn. Then models are trained to predict the impact of independent descriptive variables on the dependent churn variable. Finally, the models are evaluated and put into use for future predictions.

Table 1 provides a summary of churn prediction literature from the telecommunications domain. Telecommunications can be further divided into prepaid and postpaid services. Prepaid services, which are offered without a subscription or contract, tend to include the least amount of customer information and even what is available may not always be accurate because of sharing SIM cards. These accounts may often be shorter in duration when compared with postpaid subscription services and completely anonymous to the service provider. Therefore, attributes based on past and current account behaviour are used to make accurate predictions about future behaviour, including eventual churn.

Customer churn prediction is a supervised classification problem where customers are either churners or non-churners. To label the customers, churn must be defined for the particular setting. This may be when a contract is terminated or when an account is deactivated as in some studies (Ultsch, 2001; Wei and Chiu, 2002; Risselada *et al*, 2010). With prepaid mobile customers, there is no contract to terminate and a SIM card can remain active for up to 12 months after a customer has stopped using it, so these dates are not useful for determining when a customer churns. Some regions allow for number portability, which allows a customer to transfer their current phone number to a new provider. In this case, the date a customer requests his or her number to be ported can be used as a date of churn, as in

Table 2 Local features

Account Features	*Reload Features*	*Usage Features*	*Usage Breakdowns*
Customer ID	Number of reloads in 60 days	Number of voice contacts in 60 days	Inbound *versus* Outbound
Start Date	Reload euros in 60 days	Number of voice seconds in 60 days	Pay&Go, Residential, SME, Corporate
Plan	Date of last reload	Number of SMS contacts in 60 days	Fixed lines, International, Call Carrier
Trial Card	Card swapped in 30 days	Number of SMS sent in 60 days	Day *versus* Night
Language			Days of the Week
Handset Features			< 30 s and < 60 s

(Verbeke *et al*, 2014). If none of these are present in the date, researchers (Dasgupta *et al*, 2008; Owczarczuk, 2010; Chen *et al*, 2012) instead choose to use a period of time, between 1 and 6 months, without incoming or outgoing calls to define churn.

Separate from the issue of labelling and churn definition, the classification technique is another consideration. Many techniques have been applied to this problem including decision trees, logistic regression, support vector machines, neural networks, self-organizing maps, and genetic algorithms. Verbeke *et al* (2012) performed a benchmark study to evaluate several classification techniques for churn prediction. More recently, relational learners based on social network analysis have been introduced as alternative approaches to churn prediction (Dasgupta *et al*, 2008; Richter *et al*, 2010; Verbeke *et al*, 2014). Rather than making predictions based on a customer's attributes, relational learners view churn as a viral-like spreading through a network (Verbeke *et al*, 2014). On the basis of the papers selected for this literature review, decision trees, logistic regression, and neural networks were the most commonly used models.

Finally, the models must be evaluated. As churn is often imbalanced, with less than 10% of churners in a data set, standard measures such as accuracy are not particularly suitable to this problem. While the area under the ROC curve is often reported, this measure evaluates the predictions over the entire sample. Lift measures the percentage of highly ranked churners compared with the actual percentage of churners in the data set (Dasgupta *et al*, 2008). Since only the top-ranked customers will be contacted in the retention campaign, the more would-be churners found in that group, the more profitable a retention campaign can be. When evaluating based on other threshold metrics such as classification error and true positive rates, one should take the class imbalance into consideration. The evaluation criteria used in the literature are also indicated in Table 1.

Methodology

The data used in this study include customer information and call detail records from a major Belgian telecommunications provider. The two types of data, customer and call details, are stored separately, but an anonymized phone number can be used to link a customer from the first data set to his or her calls in the second data set. The data set includes 1.7 million prepaid (no contract) mobile customers over 6 months from May 2010 to October 2010. About 300000 were removed because they did not make or receive any calls during the 6 month observation period.

Table 3 Network features

Number A		Number of Churn Neighbours
Number B	⇒	Seconds calling churners
Start Date		Number Non-Churn Neighbours
Start Time		Seconds calling non-churners
Duration	⇒	Out-of-Network Neighbor
Call Carrier		Seconds calling out-of-network

The customer information, referred to as Local Features, includes plan data, reload amounts, handset attributes, and counts and times of various categories of calls or messages. In total, there are 111 local attributes recorded in the data set shown in Table 2. The usage features in the third column are further broken down into different categories from the fourth column to provide more finely detailed features. As is commonly the case with prepaid customers, no personal information is available. One element of this research is to make accurate predictions without this knowledge.

The call detail records, which will be transformed into Network Features shown in Table 3, include date, time, origin, and destination information about each call placed by a customer. The call information in the customer information is aggregated data, for example 9 calls to residential numbers in the last 30 days. Here, on the other hand, each record pertains to one specific call, for example, XXX called YYY on 17 June 2010 at 17:56 for 133s.

The first step in the churn prediction process is identifying the known churners. As discussed previously, defining churn can be particularly difficult in a prepaid mobile setting. While postpaid customers formally end a contract, a prepaid customer is often only formally ended after a full year of non-use. This definition, however, does not allow the company to intervene before losing the customer. Therefore, for this research, a customer is considered to have churned when they have not made or received a call for more than 30 days, based on discussions with industry experts and taking into consideration the limited amount of time covered in the data set. This is the shortest period found in extant literature for mobile customer churn. For this application the cost of misclassifying a churner is considerably higher than misclassifying a non-churner,

therefore the shorter period of time was chosen. This also allows for on-going automatic updating on a monthly basis of customers who have churned in the previous month for model checking and updated predictions in a live business setting.

The training set for this study was built using the customers and call records of the first month of observation. Customers who never made any calls during the study were removed from the data set and customers who churned during the first month were labelled as churners. The local attributes were available in a customer record data set. In order to derive the network attributes, the customer data set of approximately 1.4 million customers and the call data set comprising nearly 200 million call records were combined to form a network of customer nodes and call edges. A Java application was developed to process the data, store the network structure, and extract the network features.

Social network analysis and traditional models approach prediction from different perspectives. Traditional approaches are based on the premise that similar instances are likely to belong to the same class, and therefore use local attributes belonging to the instances. Social network analysis, on the other hand, assumes instances which are linked within a network are likely to belong to the same class, even if their individual attributes are different. One simple yet powerful concept in social network analysis is homophily, the principle that contact between similar people occurs at a higher rate than among dissimilar people (McPherson *et al*, 2001). Different ways to analyse homophily have been suggested, including relative frequency of similarity edges and dissimilarity edges and the difference between actual similarity edges and a random assignment of edges. Rhodes and Jones (2009) use the similar local attributes of nodes to predict links between them, while Zhang *et al* (2012) use the links to predict attributes of the related nodes. Both take advantage of the idea of homophily.

Compared with a network where edges are drawn randomly between any two nodes, network effects are found when more pairings occur between same-class nodes or more pairings occur between different-class nodes. One homophily test defined by Easley and Kleinberg (2010) states: 'If the fraction of cross-gender edges is significantly less than 2pq [expected cross-gender edges in a random network], then there is evidence for homophily'. In this study, $2pq = 0.3384$, while the actual cross-gender edges equals 0.1391. The difference between expected and actual cross-gender edges can be tested using standard statistical significance measures as one would test the deviation from a mean (Easley and Kleinberg, 2010). Here, using a *t*-test, the hypothesis that the values do not differ is rejected with p-value < 0.001. This indicates there is evidence for homophily, and therefore the network should provide meaningful predictive power. A small subset of the network, displaying a cluster of churners, is shown in Figure 1. White nodes are censored because they did not churn during the study. Light grey nodes represent customers who churned early in the study and dark grey nodes represent customers who churned later in the study. It can be seen that the dark grey nodes have several neighbours who had previously churned. The figure was generated using Pajek—Program for Large Network Analysis. More information about this software can be found at http://pajek.imfm.si/doku.php (accessed 11 September 2013).

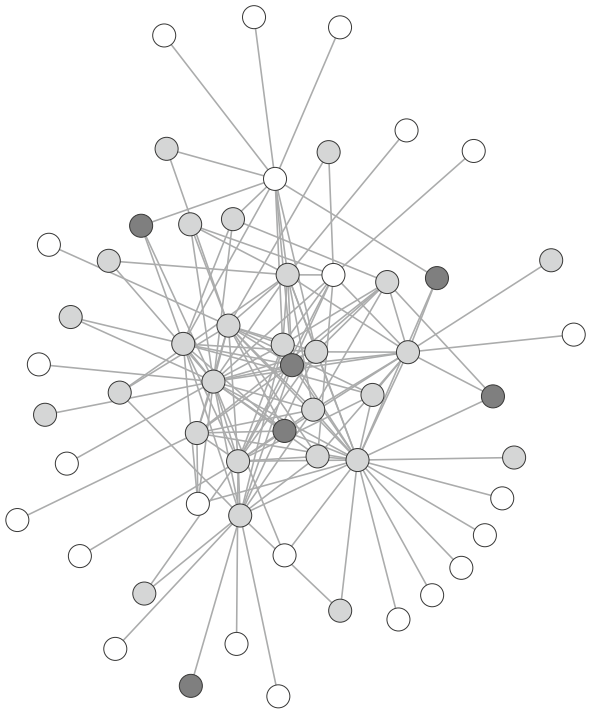

Figure 1 Cluster of churners in the network. White nodes did not churn during the study, light grey nodes churned during the first month of the study, and dark grey nodes churned later in the study.

Macskassy and Provost (2007) published an extensive study on network classification. A network is a graphical representation of data with instances as nodes and some relationship linking different nodes. Classic social networks model people and interactions, but more generally networks can model any real or representational relationship between entities. Recent publications of network analysis have been applied in many areas including nursing (Pow *et al*, 2012), behaviour adoption (Mertens *et al*, 2012), patent classification (Liu and Shih, 2010), fraud detection (Chiu *et al*, 2011), and prison system communication (Hancock and Raeside, 2009). Dasgupta *et al* (2008) applied social network analysis, specifically a diffusion model, to churn prediction in the telecommunications domain, however, with the main intent of identifying the most relevant features rather than improving accuracy.

Presenting a network as a graph allows the use of established probabilistic techniques for collective inference. The notation used here is based on that used by Macskassy and Provost (2007).

Graph $G = (V,E,C)$ where V is the set of vertices, E is the set of edges, and C is the set of class attributes $C = \{C_1, C_2, \ldots, C_p\}$. Every vertex $v_i \in V$ will have a class $c_i = C_m \in C$, however this class is not always known. Therefore, the set of all vertices can be divided into the sets of vertices with unknown and known

class attributes such that $V = V^U \cup V^K$. Edge $e_{ij} \in E$ then relates two vertices v_i and v_j with a given weight w_{ij}. The weight can be any similarity measure indicating the strength of the relationship or link between two nodes.

A Markov assumption is made to simplify the estimation of the full probability distribution of the unknown vertices: $P(c_i = C_m|G) = P(c_i = C_m|\mathcal{N}_i)$, where $\mathcal{N}_i$ is the neighbourhood of v_i, the set of vertices linked to v_i. In other words, predicting a class label only requires information about the neighbours. The first degree neighbourhood of v_i includes all vertices with an edge connecting them to v_i, while the second degree also includes the first degree neighbourhood and all the additional vertices with an edge connecting them to a vertex in the first degree neighbourhood. This can be extended to the nth degree. The results of Macskassy and Provost (2003) indicate the first degree neighbourhood is sufficient for predictions.

In the network derived for this study, each node represents a customer and one additional node represents all out-of-network destinations. Each call detail record is represented as an edge between two nodes. All contact between two customers is represented with a single undirected edge, so we do not distinguish between Customer A initiating a call to Customer B or Customer B initiating a call to Customer A. The weight of each edge is equal to the total number of seconds spent calling one another. While directed edges provide more information and may improve predictions, for this study the undirected representation was chosen to reduce computational complexity. Text messages between customers were not included in the data set and therefore are not incorporated into the social network analysis. The aggregate number of text messages for a particular customer are included in the local attributes.

Established social network analysis methods tend to begin with models learned using instance attributes, and these results are then used as prior estimates for relational learners. It is also possible to work in reverse by using social network information as attributes in a non-relational model. Featurization derives attributes from the network to be used as features in a local classifier. Lu and Getoor (2003) examine three link features, which can be extracted from networks: mode-link, count-link, and binary-link. To demonstrate these three possible derived features, refer to Figure 2, where the node of interest is grey and the neighbourhood is comprised of black (B) and white (W) nodes. The grey node has five neighbours in the black class and two neighbours in the white class. Mode-link is a single value, which represents the most common class of a node's neighbourhood; therefore, the mode-link of the grey node is B or black. Count-link is the frequency of all classes within the neighbourhood. Since the grey node has five neighbours in the black (B) class and two neighbours in the white (**W**) class, the count-link of the grey node is Black (B) 5 and White (W) 2 or $\{B,W\} = \{5,2\}$. Binary-link is a binary indicator for each class indicating whether or not it is present in the neighbourhood. Here the grey node has neighbours in both the black class and white class, so the binary-link feature is $\{B,W\} = \{1,1\}$.

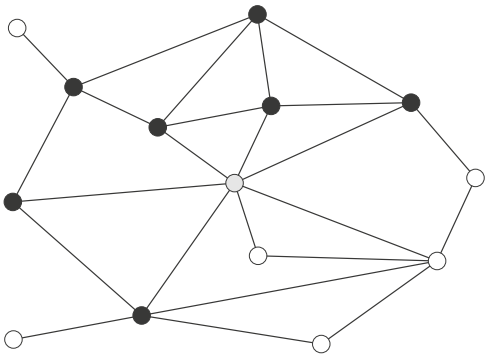

Figure 2 A sample neighbourhood, with a central grey node of interest, connected by links to neighbours of class white (**W**) and black (B).

These features can then be used in models such as support vector machines or regression models. In their work, Lu and Getoor (2003) found that building separate logistic regression models for instance variables and link features performed better than a single logistic regression model including both types of attributes. They also found that incorporating links between training and test sets is realistic and improves performance, when compared with models ignoring any links between nodes in the training and nodes in the test set (Lu and Getoor, 2003). Dasgupta *et al* (2008) identified link-based classification as an area of future research, which could bridge the gap between local attributes and social network information in classification.

The features extracted from the customer call network are displayed in Table 3, including the number of churn neighbours and total call time with churners, number of non-churn neighbours and total call time with non-churners, and also the total call time with out-of-network numbers. The number of out-of-network neighbours was not recorded. Because all out-of-network neighbours are represented by a single node in the network, this attribute is binary; 1 if the customer communicated with an out-of-network number and 0 if not.

Once all the features are available, a suitable classification modelling technique can be chosen. Survival analysis is a class of statistical models with three elements that differentiate it from other techniques: events, censoring, and time-based prediction (Kleinbaum and Klein, 2005). Events are most often defined as negative occurrences, such as loan default, and are sometimes called failures for this reason. Events may be single events, such as closing an account, or recurrent events, which can happen repeatedly like criminal recidivism. In addition, there may be multiple possible events, referred to as competing risks, such as loan default or early repayment.

Censoring occurs when something is known about the survival time, but the information is incomplete (Allison, 2010). Three types of censoring are right, left, and interval censoring, depending on what information is unknown. Right censoring is when it is known only that an event has not occurred up to a certain point in time. Left censoring is when it is known only that an event occurred before a certain point

in time. Interval censoring is when it is known that an event occurred between two distinct points in time. Right and left censoring can be seen as types of interval censoring where one end point of the interval is either 0 (left censoring) or infinity (right censoring). In survival analysis, censored information is included in the analysis, as opposed to omitting it. This allows more information, and sometimes more recent information, to be used (Im *et al*, 2012).

Finally, the quantity of interest of survival analysis is time-to-event, rather than a binary variable that would only indicate whether or not an event took place (Im *et al*, 2012). This type of prediction can allow companies to compute the profitability over a customer's lifetime (Banasik *et al*, 1999; Baesens *et al*, 2005), which may lead to a different decision about which accounts are good compared with a binary prediction. In addition, the probability of failure can be predicted consistently across many different periods of time (Im *et al*, 2012).

For these three reasons, survival analysis is suited especially well for churn prediction applications, though decision trees, logistic regression, and neural networks are the most commonly used in the literature. Churn can naturally be seen as an event as defined above, specifically a single negative event. Censoring is particularly important because churn will occur at some point in every customer trajectory. While traditional classification techniques will group customers into churners and non-churners, with survival analysis non-churners are more accurately viewed as right-censored data, in the sense that they have not yet churned but will at some later point. Time-to-event predictions allow a company to better estimate the future value of a customer, which can lead to improved decision making.

Taking these points into consideration, survival analysis, in the form of the Cox proportional hazards (PH) model, was the first modelling approach employed for this study. The Cox PH model (Cox, 1972) is a commonly used mathematical model for survival analysis. The formula for the Cox PH hazard function is:

$$h(t,\mathbf{X}) = h_0(t)\exp\left(\sum_{t=1}^{p}\beta_t X_t\right),$$

where $\mathbf{X}=(X_1,X_2,\ldots,X_p)$ is a vector of explanatory variables. The formula can be divided into two parts: the non-parametric baseline hazard $h_0(t)$ and the exponential $\exp(\sum_{t=1}^{p}\beta_t X_t)$. The baseline hazard is dependent only on time, representing the hazard function independent of the features of any instance. On the other hand, the exponential is dependent only on the instance variables and is independent of time. An extended Cox model, outside the scope of this paper, allows for time-dependent variables. The product of these two parts equals the individual's hazard at time t. The Cox PH model is semiparametric, because the baseline hazard function is not specified (Zuashkiani *et al*, 2005). Even so, it is still possible to estimate the β values and from these, the hazard ratios which measure the effect of each explanatory variable. These are estimated by maximizing the partial likelihood rather than the full likelihood (Allison, 2010). Cox (1975) has shown that standard limiting distributions for maximum likelihood estimates and tests continue to hold for partial likelihoods. It has been shown that this semiparametric model will often provide acceptable estimates compared with those that would be found if a correct parametric model was known and used (Kleinbaum and Klein, 2005).

In this study, Cox proportional hazards models were built with network attributes, local attributes, and both types of attributes combined. All models were estimated in SAS, which is accomplished with an iterative maximization process (Allison, 2010). When estimating parametric regression models, SAS uses the maximum likelihood method. Maximum likelihood estimators, particularly when using a large sample size, have been shown to be consistent, asymptotically efficient, and asymptotically normal. The probability of each observation is represented by a probability density function (pdf) $f_i(t_i)$ for uncensored observations and the survivor function $S_i(t_i)$ for censored observations. The subscript i indicates that each individual has a different function depending on their values for the covariates. The product of these functions for all observations gives the likelihood function:

$$L = \prod_{i=1}^{n}[f_i(t_i)]^{\delta_i}[S_i(t_i)]^{1-\delta_i},$$

where δ_i indicates whether an observation experiences an event or not and ensures the appropriate function (pdf or survivor) is included in the product for each individual. Depending on the parametric model being estimated, appropriate functions are substituted into the equation above in place of $f_i(t_i)$ and $S_i(t_i)$ and the expression may be transformed by taking the logarithm. Values for the parameters of these functions, represented as a vector β, are found which maximize the expression. In SAS, the Newton-Raphson algorithm is used to solve for β.

When estimating semiparametric models, like the Cox proportional hazards model used for this research, the β parameters can be estimated without specifying the baseline hazard function. As explained above, the likelihood function is the product of two parts, one dependent on the baseline hazard function and β, the other depending only on β. The partial likelihood process disregards the first part and maximizes the second part as described above to estimate β. Partial likelihoods remain consistent and asymptotically normal, but they do result in slightly higher standard errors. However, the baseline hazard function is needed in order to generate predictions for observations given their covariate values. After β is estimated by partial likelihood, the baseline hazard function can be estimated by a nonparametric maximum likelihood method. Three estimators for this maximization process are possible in (SAS Institute Inc., 2012). In this study, the Breslow (1972) method, also called empirical cumulative hazards method, was used to estimate the baseline function. Lin (2007) describes the Breslow method, its relationship to the Cox proportional hazards model, and the practical and theoretical implications of the estimator.

For model validation and evaluation, the set of customers was divided into a training set (70%) and a test set (30%). Because there are a large number of local attributes and many

with overlapping information (for example, total number of SMS, daytime SMS, nighttime SMS, Monday SMS, Tuesday SMS, etc), the feature set was reduced to remove highly correlated variables using the Wald test and insignificant variables. To avoid false positives, a Bonferroni correction was used to set an appropriate, more conservative, significance level (p-value < 0.0001). This resulted in a reduction from 111 to 39 local attributes to be used in the local model. The first five network features listed above were significant in the network model. Only the out-of-network neighbour binary variable was not significant, most likely because nearly all customers have at least one contact who is not on the same network. In the combined model, 41 features were found to be significant.

To verify the findings across other suitable classification models, logistic regression and neural networks were trained using network and local features separately. In total, 15 models are included in the comparison. Three Cox PH models, one for each feature set, were learned as these models can predict for months 1,2, and 3 in the same model. Six models each of logistic regression and neural networks were learned: two feature sets over 3 months. After the models were learned, they were used to predict the churn time of the customers in the test set. In Cox proportional hazards models, the output is a probability of event, in this case churn, for each time interval. The results below are based on 1 month time intervals.

In order to evaluate the resulting models, the ROC, AUC, and Lift were calculated and compared. As the ultimate goal of churn prediction is to increase profits, this study also uses the expected maximum profit criterion (Verbraken *et al*, 2013), an extension of the maximum profit criterion introduced by Verbeke *et al* (2012). These measures are based on a cost/benefit analysis framework aligned to the goal of profit maximization. They applied the framework to the specific area of customer churn prediction and developed the Expected Maximum Profit Measure for Customer Churn, (EMP^{ccp}). A probabilistic method is used for the expected measure as it is assumed that not all parameters are known. Below is their definition of the expected maximum profit measure,

$$EMP = \int_{b_0} \int_{c_1} \int_{c^*} P(T(\theta); b_0, c_1, c^*) \cdot w(b_0, c_1, c^*) dc^* dc_1 db_0$$

with T the optimal threshold, b_0 the incremental benefit of true positive, c_1 the incremental cost of a false negative, c^* the cost of the action imposed on the instance, and $w(b_0,c_1,c^*)$ the joint probability density of the classification costs and benefits. Here, b_0, c_1, and c^* can take any positive real value. It is assumed that $b_0 > c^*$, because it only makes sense to undertake a cost when the expected benefit is greater. The expected maximum profit is calculated by integrating over all possible benefits and costs.

Choosing appropriate values for these parameters is not always straightforward as they will vary by industry and company. A specific company wishing to use this measure to evaluate their churn models should calculate estimates for customer lifetime value, incentive cost, and contact cost. The estimates established by Verbraken *et al* (2013) are CLV = €200, incentive cost = €10, and contact cost = €1. These same estimates are used in this study after consulting with the business users involved in this research. The probability of a churner accepting the incentive is more difficult to determine. This uncertainty is introduced via a probability parameter. These are then substituted into a profit function defined by Neslin *et al* (2006) for the overall profit of a retention campaign:

$$Profit = N\eta[(\gamma CLV + d(1-\gamma))\pi_0\lambda - d - f] - A,$$

where N is the total number of customers, η is the fraction of customers who will receive the incentive, γ is the success rate of the incentive, d is the incentive cost, f is the contacting cost, and A is the fixed administrative cost. In addition, λ is the percentage of churners in η divided by the base churn rate π_0.

This profit function is then incorporated into the general EMP measure to find the EMP for customer churn as well as the ideal fraction of customers to contact.

$$EMP^{ccp} = \int_{\gamma} P_C(T(\gamma); \gamma, CLV, \delta, \theta) \cdot h(\gamma) d\gamma,$$

with T, the optimal threshold for a given γ and $h(\gamma)$ the probability density function for γ. As stated above, a probability distribution is used because there is uncertainty about the acceptance rate of potential churners; a β distribution is used for flexibility. On the basis of the range for acceptance rates assumed by Neslin *et al* (2006), the parameters for the distribution proposed by Verbraken *et al* (2013) are based on an expected value of 30% and standard deviation of 10%. While theoretically the general *EMP* is integrated over all possible costs and benefits, the assumptions made for the cost and benefit parameters limit the integration of EMP^{ccp} to a portion where the benefits of contacting churners is greater than the cost of contacting the targeted customers.

The calculation of the ideal fraction of customers to target with the retention campaign is an additional benefit offered by using the EMP approach. Companies can make decisions with advance expectations about the appropriate threshold, fraction of customers to contact, and expected profitability of the campaign. The fraction of customers to target in order to maximize the expected profit is:

$$\eta emp = \int_{\gamma} [\pi_0 F_0(T(\gamma)) + \pi_1 F_1(T(\gamma))] \cdot h(\gamma) d\gamma,$$

with T, the optimal threshold as above.

Findings

In evaluating the significant local attributes, most hazard ratios are just above or just below 1, indicating a small increase or decrease in probability to churn. The attribute associated with the greatest increase of churn risk was a binary indicator if the account had a trial card, with an increased risk of 25.2%. On the other hand, having an additional 'friend' on the same call carrier

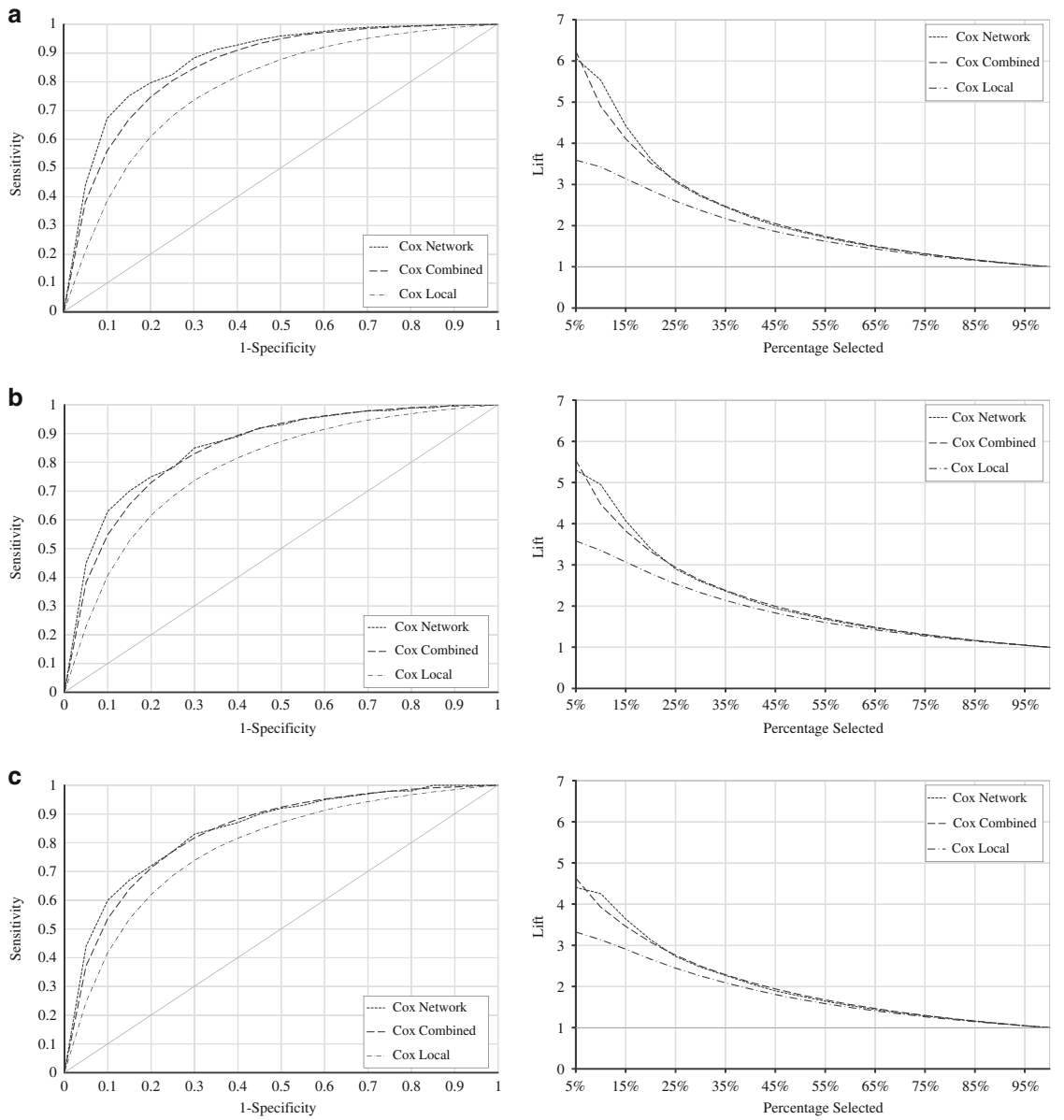

Figure 3 Comparison of Models: ROC Plots and Lift Charts.
(a) Month 1 ROC Plot and Lift Chart.
(b) Month 2 ROC Plot and Lift Chart.
(c) Month 3 ROC Plot and Lift Chart.

decreases the risk of churn by 26.6%. The network attributes are much more striking. Every additional churn neighbor increases the probability of churn 2.85 times. Conversely, for each non-churn neighbor the probability of churn decreases by 35.9%.

Over the 3 months and 15 models, those trained on network features outperform those trained on local features. There is less difference in the performance among model types. All local only models are nearly even with each other, but well below the other three models. The combined model is between the network only and local only models. Because all network models give similar results and all local models give similar results, the ROC curves are plotted for just the Cox PH models over 3 months in Figure 3. A plot of all models would show a cluster of network models and another cluster of local models, but the individual curves are difficult to identify, in that case. ROC curves display graphically the true positive rate *versus* the false positive rate. The closer an ROC curve is to the point (0,1), the better the model is (DeLong *et al*, 1988). Because the output of the models is represented as a probability, a cut-off score or threshold is used to divide the predicted churners from the predicted non-churners. As this cut-off score changes, the associated true/false positive rates also vary. Each point on the ROC curve corresponds to a possible cut-off score.

Table 4 Model evaluation results

Model	*Features*	*AUC*	10% *Lift*	EMP^{ccp}	*η emp (%)*
Month 1					
Logistic Regression Network	6	**0.8836**	5.5000	0.1404	3.81
Neural Network Network	6	0.8766	**6.2847**	0.1341	3.98
Cox PH Network	5	0.8735	5.5350	**0.1408**	4.05
Cox PH Combined	41	0.8691	4.8980	0.1154	3.13
Logistic Regression Local	29	0.7872	3.5210	0.0004	0.05
Neural Network Local	38	0.7774	3.4427	0.0043	0.200
Cox PH Local	39	0.7792	3.4280	0.1395	5.73
Month 2					
Logistic Regression Network	6	0.8601	4.9370	0.9206	9.62
Neural Network Network	6	**0.8613**	**5.5395**	**0.9385**	10.04
Cox PH Network	5	0.8543	4.9518	0.9383	9.92
Cox PH Combined	41	0.8553	4.4745	0.7815	9.96
Logistic Regression Local	29	0.7850	3.3775	0.2773	8.12
Neural Network Local	38	0.7814	3.3402	0.2392	7.74
Cox PH Local	39	0.7794	3.3534	0.7739	14.65
Month 3					
Logistic Regression Network	6	0.8436	4.2523	1.9045	13.87
Neural Network Network	6	**0.8480**	**4.9502**	1.9576	14.55
Cox PH Network	5	0.8425	4.2569	1.9378	13.78
Cox PH Combined	41	0.8462	3.9274	1.7833	16.03
Logistic Regression Local	29	0.7843	3.1120	1.1672	17.24
Neural Network Local	38	0.7877	3.2603	1.2067	18.10
Cox PH Local	39	0.7794	3.1349	**2.0890**	27.83

Fifteen models were trained using either local instance variables, network features, or a combination of both. They were evaluated for each of 3 months, according to the area under the ROC curve, 10% lift, the expected maximum profit for churn prediction, and the ideal percentage of customers who should be contacted according to that EMP estimate. Because a churner in the first month will still be considered a churner for the future months, models can only be compared within the same month. For each month, the greatest AUC, the highest lift, and the greatest EMP are highlighted in bold.

The ROC curves show the improved prediction power of network attributes. In particular, the leftmost part of the curve is the area of interest in a churn campaign, as this represents only the most likely churners. The sharper incline of the network models indicates a better selection of those likely to churn and therefore those who should be contacted as part of a retention campaign. Recall that the Cox PH model is a single model capable of making predictions for each time period, while a separate logistic regression or neural network model is trained for each time interval.

In order to compare the models statistically, the AUC (area under the ROC curve) is included in Table 4. The AUC value is interpreted as the probability that a randomly chosen positive case will be ranked higher than a randomly chosen negative case. It can be estimated by calculating the related Mann-Whitney statistic (Bamber, 1975). The AUC values are included in Table 4. As seen in the ROC curves, the AUC does not show a great difference between model types. However, regardless of the model used, the network attributes do provide significant improvement in predictive power over the local attributes. This corroborates the findings of the paper of Benoit and den Poel (2012), which investigated the impact of adding network attributes to churn prediction models in the financial services sector and found significant improvement when incorporating social network features.

While the ROC plots and AUC values show the performance over the whole range, in churn prediction ranking at the top is most important. Lift charts for the Cox PH models are provided in Figure 3 to better evaluate the differences in the highest rankings. As with the ROC curves, if all lift curves were plotted together, they would form two distinct groups: those including network attributes and those with only local attributes. When limited to only the top 10% most likely churners, the models which include network features show a lift of more than 5.5 compared with the baseline for Month 1 predictions, while the local only models have a lift less than 3.5. It can be observed that the differences between the models lessen over time. The lift values of the local only models do not change much between Month 1 and Month 3, but the lift values of the models including network features decrease each month. This suggests that the network features provide more information on churn in the near future, while local features may predict longer term behaviour.

As profit is the main driver behind churn prediction, these models have also been evaluated using the expected maximum profit measure for churn prediction. Table 4 shows the EMP^{ccp} in euros per customer for the classification models from this study, in addition to their AUC values. The profit measure gives conflicting results over the 3 months. In the first 2 months, the network features outperform the local features, but interestingly the Cox PH local model has the highest expected profit

for the third month. Because the profit estimates are given as euros-per-customer and there were 1.4 million customers in this study, a relatively small increase in EMP represents a very large impact on overall profitability.

In Month 1, the expected maximum profit of any network model as well as the the Cox PH local model is €0.13 per customer greater than the expected profit of the neural network and logistic regression local models. Very little profit is possible under the assumed parameters when using the neural network local model or logistic regression local model for ranking likely churners. In Month 2, the network models result in 15 cents more per customer compared with the local Cox PH model, a difference of about €210000 for the campaign. As in the first month, the neural network local model and logistic regression local model result in a much less lucrative campaign, 50 to 70 cents per customer, or €979000 overall, less than the network models. In Month 3, the Cox regression local model has the advantage over the network models; however to obtain this profitability, twice as many customers should be offered the incentive, which may be a deterrent if resources are limited. Again, the neural network local model and logistic regression local models result in much lower expected profits.

When comparing the profit-based results, it is important to note that the EMP^{ccp} increases each month for all models because of the cumulative way of evaluating churners. Any customer who has churned during a previous month is considered a churner for the current month. Therefore, the EMP^{ccp} can only be compared within a single month across models.

Conclusion

Summary

This study compared the usage of local customer and relational network attributes in churn prediction. Previous literature, though limited, suggests that social networks can add information about a customers' likelihood to churn. The challenge is to access and use this network data in an efficient way. This study contributes a business-oriented approach, implemented in Java and analysed using SAS, which allows a company to use data generated in 1 month to make accurate—and more importantly profitable—predictions about churn in the coming months and precisely plan incentives to intervene.

The findings demonstrate how incorporating social network features into a churn prediction model can enhance overall prediction results and improve the profitability of a retention campaign. In other words, this study indicates the presence of valuable information in customers' social networks. Accessing and using this information can greatly impact the profitability of a campaign to intervene with potential churners. Currently, churn prediction is made using customer data, but according to the analysis here based on lift, AUC, and EMP, using the social network results in superior classification and profitability. By using Cox proportional hazards models, the likelihood of churn for each individual customer can be determined for each future time segment based on a single model. This allows for highly targeted marketing campaigns, which can be used to intervene with a smaller subset of customers in a timely manner, resulting in reduced costs and higher potential for rewards, estimated at more than 15 cents per customer when comparing the Month 2 predictions of local models and network models as shown in Table 4.

With regards to profitability, the models using network features offer a greater expected maximum profitability when used to make predictions within 2 months, though the AUC is higher in all 3 months. This suggests that network information is more time sensitive than local attributes. The EMP^{ccp} measure also indicates that the intervention group should be a smaller subset of customers for maximum profitability using the network model, indicating a more targeted marketing campaign is possible.

Limitations and future works

This study is based on a one company's data set including 6 months of data. In this case, the network's high degree of homophily led to strong network effects and improved resulting accuracy. Additional data sets, including different companies or longer time frames could further validate the findings.

The data used for this research were collected in 2010, but mobile communications are continually evolving. As mobile data usage continues to supplement and even replace traditional voice and text communications, it would be interesting to extend this research to include different types of networks beyond the call graphs used in this study.

Acknowledgements—This research was made possible with support of the Odysseus program (Grant B.0915.09) and Grant G.0816.12N of the Fund for Scientific Research Flanders (FWO).

References

Abbasimehr H, Setak M and Soroor J (2013). A framework for identification of high-value customers by including social network based variables for churn prediction using neuro-fuzzy techniques. *International Journal of Production Research* **51**(4): 1279–1294.

Aksoy L, Buoye A, Aksoy P, Lariviere B and Keningham T (2013). A cross-national investigation of the satisfaction and loyalty linkage for mobile telecommunications services across eight countries. *Journal of Interactive Marketing* **27**(1): 74–82.

Allison PD (2010). *Survival Analysis Using SAS: A Practical Guide*. 2nd edn, SAS Institute Inc.: Cary, NC.

Baesens B, Van Gestel T, Stepanova M, Van den Poel D and Vanthienen J (2005). Neural network survival analysis for personal loan data. *Journal of Operational Research Society* **56**(9): 1089–1098.

Bamber D (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* **12**(4): 387–415.

Banasik J, Crook JN and Thomas LC (1999). Not if but when will borrowers default. *Journal of Operational Research Society* **50**(12): 1185–1190.

Benoit D and den Poel D Van (2012). Improving customer retention in financial services using kinship network information. *Expert Systems with Applications* **39**(13): 11435–11442.

Bersen A, Smith S and Thearling K (2000). *Building Data Mining Applications for CRM*. McGraw-Hill: New York.

Breslow NE (1972). Contribution to the discussion on the paper by DR Cox, regression and life tables. *Journal of the Royal Statistical Society* **34**(2): 216–217.

Chen Z-Y, Fan Z-P and Sun M (2012). A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data. *European Journal of Operational Research* **223**(2): 461–472.

Chiu C, Ku Y, Lie T and Chen Y (2011). Internet auction fraud detection using social network analysis and classification tree approaches. *International Journal of Electronic Commerce* **15**(3): 123–147.

Cox DR (1972). Regression models and life-tables. *Journal of the Royal Statistical Society* **34**(2): 187–220.

Cox DR (1975). Partial likelihood. *Biometrika* **62**(2): 269–276.

Dasgupta K *et al* (2008). Social ties and their relevance to churn in mobile telecom networks. In: *EDBT '08 Proceedings of the 11th International Conference on Extending Database Technology: Advances in Database Technology*, pp 668–677, Nantes, France, March. ACM.

de Bock KW and den Poel D Van (2011). An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction. *Expert Systems with Applications* **38**(10): 12293–12301.

DeLong E, DeLong D and Clarke-Pearson D (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *International Biometric Society* **44**(3): 837–845.

Easley D and Kleinberg J (2010). *Networks, Crowds, and Markets*. Cambridge University Press: Cambridge.

Hancock PG and Raeside R (2009). Analysing communication in a complex service process: An application of social network analysis in the scottish prison service. *Journal of Operational Research Society* **61**(2): 265–274.

Huang B, Kechadi MT and Buckley B (2012). Customer churn prediction in telecommunications. *Expert Systems with Applications* **39**(1): 1414–1425.

Im J-K, Apley DW, Qi C and Shan X (2012). A timedependent proportional hazards survival model for credit risk analysis. *Journal of Operational Research Society* **63**(3): 306–321.

Kim H-S and Yoon C-H (2004). Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. *Telecommunications Policy* **28**(9–10): 751–765.

Kleinbaum DG and Klein M (2005). *Survival Analysis: A Self-Learning Text*. Springer: New York.

Lam SY, Shankar V, Erramilli MK and Murthy B (2009). Customer value, satisfaction, loyalty, and switching costs: An illustration from a business-to-business service context. *Journal of the Academy of Marketing Science* **32**(3): 293–311.

Lima E, Mues C and Baesens B (2009). Domain knowledge integration in data mining using decision tables: Case studies in churn prediction. *Journal of Operational Research Society* **60**(8): 1096–1106.

Lin DY (2007). On the Breslow estimator. *Lifetime Data Analysis* **13**(4): 471–480.

Liu D-R and Shih M-J (2010). Hybrid-patent classification based on patent-network analysis. *Journal of the American Society for Information Science and Technology* **62**(2): 246–256.

Lu Q and Getoor L (2003). Link-based classification using labeled and unlabeled data. In: *Proceedings of the ICML Workshop on The Continuum from Labeled to Unlabeled Data*, ICML: Washington DC.

Macskassy SA and Provost F (2003). A simple relational classifier. In: *Proceedings of the Second Workshop on Multi-Relational Data Mining (MRDM-2003) at KDD-2003*. ACM: New York, NY, pp 64–76.

Macskassy SA and Provost F (2007). Classification in networked data: A toolkit and a univariate case study. *Journal of Machine Learning Research* **8**(2): 935–983.

McPherson M, Smith-Lovin L and Cook JM (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology* **27**(1): 415–444.

Mertens F, Saint-Charles J and Mergler D (2012). Social communication network analysis of the role of participatory research in the adoption of new fish consumption behaviors. *Social Science and Medicine* **75**(4): 643–650.

Neslin SA, Gupta S, Kamakura W, Lu J and Mason CH (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research* **43**(2): 204–211.

Nitzan I and Libai B (2011). Social effects on customer retention. *Journal of Marketing* **75**(6): 24–38.

Owczarczuk M (2010). Churn models for prepaid customers in the cellular telecommunication industry using large data marts. *Expert Systems with Applications* **37**(6): 4710–4712.

Pendharkar PC (2009). Genetic algorithm based neural network approaches for predicting churn in cellular wireless network services. *Expert Systems with Applications* **36**(3): 6714–6720.

Pow J, Gayen K, Elliott L and Raeside R (2012). Understanding complex interactions using social network analysis. *Journal of Clinical Nursing* **21**(19–20): 2772–2779.

Rhodes CJ and Jones P (2009). Inferring missing links in partially observed social networks. *Journal of Operational Research Society* **60**(10): 1373–1383.

Richter Y, Yom-Tov E and Slonim N (2010). Predicting customer churn in mobile networks through analysis of social groups. In: *Proceedings of the 2010 SIAM International Conference on Data Mining*, SIAM: Columbus, OH, pp 732–741.

Risselada H, Verhoef P and Bijmolt T (2010). Staying power of churn prediction models. *Journal of Interactive Marketing* **24**(3): 198–208.

Rosenberg LJ and Czepiel JA (1984). A marketing approach for customer retention. *Journal of Consumer Marketing* **1**(2): 45–51.

SAS Institute (2012). The PHREG Procedure. In *SAS/ STAT*R_ *12.1 User's Guide*, SAS Institute: Cary, NC, pp 5541–5769.

Ultsch A (2001). Emergent self-organising feature maps used for prediction and prevention of churn in mobile phone markets. *Journal of Targeting, Measurement, and Analysis for Marketing* **10**(4): 314–324.

Verbeke W, Dejaeger K, Martens D, Hur J and Baesens B (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research* **218**(1): 211–229.

Verbeke W, Martens D and Baesens B (2014). Social network analysis for customer churn prediction. *Applied Soft Computing* **14**(Part C): 431–446.

Verbeke W, Martens D, Mues C and Baesens B (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications* **38**(3): 2354–2364.

Verbraken T, Verbeke W and Baesens B (2013). A novel profit maximizing metric for measuring classification performance of customer churn prediction models. *IEEE Transactions on Knowledge and Data Engineering* **25**(5): 961–973.

Wei C-P and Chiu I-T (2002). Turning telecommunications call details to churn prediction: A data mining approach. *Expert Systems with Applications* **23**(2): 103–112.

Wong K.K.-K. (2011). Using Cox regression to model customer time to churn in the wireless telecommunications industry. *Journal of Targeting, Measurement, and Analysis for Marketing* **19**(1): 37–43.

Zhang X, Zhu J, Xu S and Wan Y (2012). Predicting customer churn through interpersonal influence. *Knowledge-Based Systems* **28**(1): 97–104.

Zuashkiani A, Banjevic D and AKS Jardine (2005). Estimating parameters of proportional hazards model based on expert knowledge and statistical data. *Journal of Operational Research Society* **60**(12): 1621–1636.

Received 14 October 2013;
accepted 1 February 2016