A Comparative Study of Social Network Classifiers for Predicting Churn in the Telecommunication Industry

María Óskarsdóttir*, Cristián Bravo†, Wouter Verbeke‡, Carlos Sarraute§, Bart Baesens*† and Jan Vanthienen*

*Faculty of Economics and Business, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium,
Email: {maria.oskarsdottir, bart.baesens, jan.vanthienen}@kuleuven.be
†Southampton Business School, University of Southampton, United Kingdom, Email: c.bravo@soton.ac.uk
‡Faculty of Economic and Social Sciences and Solvay Business School, Vrije Universiteit Brussel, Belgium,
Email: wouter.verbeke@vub.ac.be
§Grandata Labs, Argentina, Email: charles@grandata.com

Abstract—**Relational learning in networked data has been shown to be effective in a number of studies. Relational learners, composed of relational classifiers and collective inference methods, enable the inference of nodes in a network given the existence and strength of links to other nodes. These methods have been adapted to predict customer churn in telecommunication companies showing that incorporating them may give more accurate predictions. In this research, the performance of a variety of relational learners is compared by applying them to a number of CDR datasets originating from the telecommunication industry, with the goal to rank them as a whole and investigate the effects of relational classifiers and collective inference methods separately. Our results show that collective inference methods do not improve the performance of relational classifiers and the best performing relational classifier is the network-only link-based classifier, which builds a logistic model using link-based measures for the nodes in the network.**

I. INTRODUCTION

Customer churn prediction (CCP) in the telecommunication industry (telco) has been intensively researched during the last decade. The market is competitive for the companies since, while it is relatively easy for unhappy customers to change providers, acquiring new customers is more expensive than retaining current ones and happy customers are more likely to attract others [1], [2]. In addition, the companies gather an abundance of data about their customers, such as demographic information, usage behaviour and call detail records (CDR) which, when analysed in the correct way, can provide valuable information and enhance the likelihood of the company to thrive in this fiercely competitive market.

In the last years, various studies have confirmed that incorporating network effects in CCP models can improve their performance greatly. One way to model network effects, is by applying propagation algorithms to call networks and thus simulate how churners might influence other users in the network. The propagation algorithms produce a score for each customer, that can either be used as an attribute in classical binary classifiers or seen as that person's probability of churning. The call networks are built by aggregating CDR datasets to create a social network of customers who are connected if they have a relationship, which in this case means that they have used their cell phones to connect. Based on the network, information flow between the customers through the strength of the connections can be simulated and used to make inferences about the customers characteristics, such as the propensity to churn. This is called network learning [3]. Various propagation methods for network learning exist. Together we refer to them as Relational Learners (RL), since they are used to learn from relationships in a network. They can be separated into two groups based one their purpose: Relational Classifiers (RC) and Collective Inference methods (CI).

In this study we expand on NetKit, a relational learning framework, presented by Macskassy and Provost [3] which was later adapted for predicting churn in telcos [4]. In the latter study, four relational classifiers and five collective inference methods were applied to call networks to investigate different ways of incorporating relational learners and whether they improve predictions. Here we focus on one of those ways, namely using the relational learners as churn prediction models themselves, thus interpreting the scores as probabilities of churn. We have collected seven distinct CDR datasets from across the world, to which we applied 24 different combinations of RC and CI. The goal of the study is threefold. Firstly, we would like to know if there is a relational learner which outperforms the others. Secondly, by looking at the relational classifiers themselves, we seek to discover if they have an internal ranking in terms of performance. Finally, we investigate the same for the collective inference methods to answer the ultimate questions of whether combining them with relational classifiers improves the performance of customer churn prediction models in telco.

Our contributions are the following:
- There is a group of relational learners which outperform the rest. Most of them use the network-only link-based classifier.
- The worst performing learners are usually the same for all performance measures. These are the learners that apply the iterative classification collective inference method.

- In the case of customer churn prediction in telco, there is no added benefit of using collective inference methods together with relational classifiers, which perform better when used on their own.

The rest of this paper is organized as follows. In section II we present related work regarding churn prediction in telco as well as inferencing in networked data. After that, we describe the methodology used in this research. We start with a discussion about networked data followed by a description of the algorithms used in the experiments. Next, we provide a short description of the measures used to evaluate the performance of the relational learners. We briefly describe the datasets we used and then explain our experimental setup. Subsequently we present the results of our experiments. Finally, the paper concludes with future work.

II. RELATED WORK

As a research domain, customer churn prediction modelling is already well established. It is a classification problem which has been studied intensely in various domains where the relationship with current customers is seen as a valuable asset for the company. Customer churn prediction has been applied in the banking sector [5], [6], [7], by insurance companies [8], [9], internet service providers [10] and in the telecommunication industry, which is the industry on which we focus here. We refer to [11] for an overview of commonly used classification techniques for churn prediction in telco and a benchmarking study of those techniques.

The benchmarking study [11] as well as many other papers mostly rely on local variables, meaning that no social behaviour was incorporated or they are not explicit about social effects in their models. However, social variables, such as social interaction and calling behaviour, are important factors when it comes to churn in telco. In this context, *local variables* include demographic and usage features for the entities in the dataset, without any information about connections to other entities. In contrast, *social variables* contain some information about the links to other entities [12]. In recent studies where social effects have been incorporated in the models it has been shown that the model performance improves greatly. In some studies the datasets have been enriched with network variables, such as degree, transitivity and centrality, before building a model using binary classifiers [13], [14], [15], [16]. Including these kinds of variables in the dataset adds valuable information that is different from the local variables, and thus better models are produced.

Other studies exploit some kind of propagation algorithms that spread 'churn influence' throughout the network to simulate how churners might possibly affect the people they are connected with. The spreading activation algorithm [17], which is often compared to the word-of-mouth effect, has successfully been used for this purpose both on its own [17], [18] and to produce scores that are then used as variables in non-relational classifiers, such as logistic regression and decision trees [19], [20]. Lu and Getoor [21] introduced a method called the network-only link-based classifier, which learns a logistic regression model for the nodes in the network

using link-based network variables. In other words, each node can be classified using only the connections to other nodes. In addition, these two modelling approaches have been combined in various ways with one study showing that the best combination model is achieved when a binary classifier is used to build a new model, using the scores resulting from non-relational classifiers and relational learners (in this case spreading activation) as variables [18].

The increase in predictive power of models that take into account social network effects can partly be explained by homophily and social influence. Homophily is used to describe peoples' tendency to associate with those that are similar to them [22]. Studies have shown that there is significant difference in the phone usage of different genders and age groups, and the usage patterns can be used to predict the demographic features [23], [24]. Additionally, homophily can be used to predict similarities between people who interact frequently or to predict interactions between people who behave in a similar way [13], [25]. Social influence, on the other hand, happens when a person's behaviour is influenced by others around them [26].

Finally, in addition to CCP in telco, social network analysis has also been shown to give improved results in other areas, such as social network online games [27] and social security and credit card fraud [28], [29].

III. METHODOLOGY

A. Networks

Networks consist of nodes, which may represent real life entities such as people, and the connections or relationships between them, the edges [30]. Weighted and labelled networks are often represented by the triple

$$\mathcal{G} = (\mathbf{V}, \mathbf{E}, \mathcal{X}) \quad (1)$$

where

$$\mathcal{X} = \{c_1, \ldots, c_m\} \quad (2)$$

is a set of m classes. The first component in the triple $\mathbf{V}$ consists of a vector of $n \in \mathbb{N}$ vertices $\mathcal{V} = \{v_1, \ldots v_n\}$ and a vector of labels $\mathcal{L} = \{l_1, \ldots, l_n\}$ where each $l_i \in \mathcal{X}$ represents the class of the node. Similarly, the edge part of the graph $\mathbf{E}$ is composed of two parts, edges and weights. The edges $\mathcal{E}$, are a set of two-subsets of $\mathcal{V}$, where the edge $e_{ij} \in \mathcal{E}$ means that there exists a connection between nodes v_i and v_j. The second component of $\mathbf{E}$ are the weights, which represent the strength of the connection between two edges. As such $w_{i,j} \in \mathbb{R}_+$ and we further assume that the networks are undirected, that is

$$w_{ij} = w_{ji}, \quad \forall \, i, j \in \{1, \ldots, n\}. \quad (3)$$

For the purpose of this study, we suppose that not all of the class labels are known. Therefore we make a distinction between the set of unknown $\mathcal{V}^U$ and know $\mathcal{V}^K$ nodes. Lastly, we define the neighborhood of a node v_i as being the collection of the node itself and all nodes that are directly linked to it or

$$\mathcal{N}_i = \{v_i\} \cup \{v_j \in \mathcal{V} \mid w_{ij} \neq 0\}. \quad (4)$$

There are various ways to define the strength of links between two nodes in a network. In the case of call networks, indicating
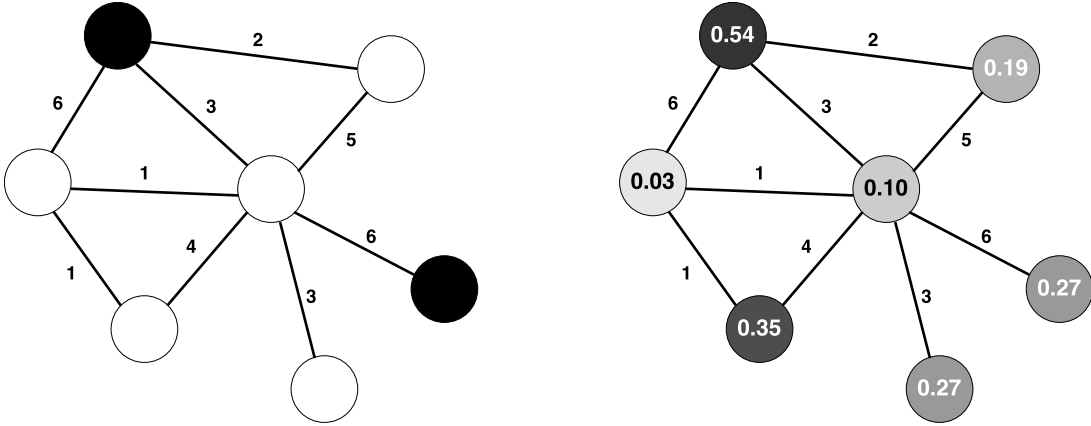
Fig. 1. The figure shows an example of an application of a relational learner. The figure on the left, displays a graph with seven customers, of which two have churned (black) and five have not churned (white). The figure on the right shows the same network after the RL has been applied. Each customer now has a score or probability of churning.

whether or not two customers had a relationship during a specific time period will result in a binary network. Other, more advanced possibilities, include keeping track of how much time customers spend talking to each other or how often they talk and send messages during a certain period of time. Representing the weights in different ways will result in different insights. In addition, the links may have different weights depending on when they were made. In that way, recent activity between two customers can be given more importance than older activity. To model this in the network, the weights at time t, $w_{ij,t}$ can be exponentially weighted in time by

$$(w_{ij})_t = e^{-\gamma t} w_{ij,t} \tag{5}$$

where γ is the decay constant. The final weights are obtained by aggregating all $(w_{ij})_t$ for the whole time period. Weighted networks have been successfully used in credit card and social insurance fraud detection [28], [29].

B. Learning in Networks

Networks representing relationships between people can be exploited for network learning, thus simulating social influence. In CCP this is particularly interesting since it is possible to estimate how churners affect the non-churners and subsequently how 'churn-influence' spreads through the network. Relational learners exploit the information flow between interlinked entities in a network. Assuming we have available information describing the characteristics of the nodes, the links between them can be used to infer these characteristics if they happen to be unknown for some nodes. The framework and toolkit NetKit was developed with this in mind [3]. It can be applied to a range of networked data, thus combining various methods to make inferences about interlinked nodes with unknown labels in a network.

In their paper, Verbeke et al. [4] adjusted the framework of NetKit to fit the requirements of customer churn prediction in telco. To make it applicable to the more specific problem setting, a few changes were made. Firstly, churn prediction is

a binary classification problem, meaning that there are only two classes which results in

$$\mathcal{X} = \{c_0, c_1\}, \tag{6}$$

where c_0 and c_1 represent non-churners and churners, respectively. Secondly, some of the collective inference methods in [3] were too slow for the large call networks, and were therefore adjusted to run pseudo-simultaneously instead of sequentially. Finally, Verbeke et al. [4] introduced the aspect of time in CCP to NetKit. In NetKit, it is always assumed that only some of the nodes have unknown labels and the rest have known labels which are used to infer the unknown ones. In churn prediction, on the other hand, the aim is to make predictions into the future, where all labels are unknown. To implement this, Verbeke et al. suggested training the algorithms at a specific time, t, where all labels are known, and use the resulting scores as the estimated labels at time $t+1$. This means that $\mathcal{V}^K = \emptyset$ at time $t+1$ before the analyses are started. This is depicted in Fig. 1 which shows a small call network before and after a relational learner is applied. The figure on the left shows the network at time t where all labels are known and two people have churned (black nodes). The figure on the right shows the same network at time $t+1$ when a relational learner has been used to simulated the propagation of 'churn-influence' resulting in a score or churn probability for each node.

In this research, relational learners are split into two groups: Relational Classifiers and Collective Inference methods.

Relational Classifiers (RC) are the methods which infer class labels for each node in a network based on the strength of links to other nodes and the labels of those nodes. They perform a single, local operation going from node to node until all have been classified. Different methods exist. The ones that were used in this study are discussed in table I.

When going through the network in this manner, it is easy to see that the classification might not be very stable. Once the first node has been classified, its class label is used to infer the class label for the second node, which in turn might change, which could again have an effect on the first node. When

TABLE I
RELATIONAL CLASSIFIERS

Abbreviation	Description
WVRN	The weighted vote relational neighbor classifier infers a score based on the weighted labels of the connected nodes.
CDRN	The class distribution relational neighbor classifier assigns a label by looking at the distribution of classes of connected nodes [31].
NLB	The network-only link-based classifier determines a score by learning a logistic regression model using the link-based measures of the nodes. [21].
SPA RC	The spreading activation relational classifier is the classification part of the spreading activation algorithm. It computes scores by looking at weights and labels of nodes connected to neighboring nodes [17].

TABLE II
COLLECTIVE INFERENCE METHODS

Abbreviation	Description
GS	Gibbs sampling is well known in image retrieval. It is an iterative method which applies an RC to the network, first for a burn-in period without keeping track of class inference and subsequently applying it and adding up the scores until all iterations are finished. In each iteration, a class label is sampled from the resulting vector of the RC. At the end the scores are normalized to get final scores [34].
IC	Iterative classification iteratively applies a relational classifier to the network, assigning the label which most often occurs in the resulting vector of estimates to each node [21].
RL	Relaxation labelling simultaneously infers scores for each node in the network by applying a relational classifier in each iteration and using the resulting estimates as input for the next iteration. [35]
RL AS	Relaxation labelling with simulated annealing is an extension of RL with an added annealing term. It is related to the PageRank algorithm [36].
SPA CI	The spreading activation collective inference method is the CI part of the SPA method. It initializes scores the same way as the gibbs sampler and then iteratively applies RC to the network and updates the scores. [17].

applying the RC a single time, this effect is not captured. To do so, collective inference methods are used to regulate the inferencing process.

Collective Inference methods (CI) are procedures which infer class labels for the nodes in a network while taking into account how the inferred labels affect each other. They decide in which order the nodes are labelled and how a final label is determined. They have been shown to improve the performance of relational classifiers in genomes and bibliographic networks [32], [33]. Table II documents the CI used in this study. In general the CIs work in a very similar way performing two operations iteratively until some terminating requirement is reached. First, a relational classifier is applied to each node in the network and then the score of each node is updated using the results of the RC.

When CIs are applied, the resulting scores have a tendency to level out and as a result, they don't have much variation. That is, when the methods repeatedly classify the nodes, a smoothing effect of the churn influence occurs, resulting in very little distinction of churners and non churners. This was verified in the early stages of this study, showing that for some CIs the variation in scores decreases very quickly. As a result, early stopping was implemented in all CI algorithms. The early stopping is similar to the mechanism which was already implemented for the spreading activation method. The criterion is that as soon as the variation in the scores is less than some threshold, the inferencing stops. The threshold was determined after a sensitivity analysis was performed on part of the data.

Each of the four relational classifiers can be combined with one of the five collective inferencing methods or applied on its own, which results in a total of 24 combinations of methods.

C. Performance Measures

To evaluate the performance of the models, lift, AUC and the H-measure will be used. Firstly, we look at the lift measure [37], which represents how much better a model is at identifying churners than one would find in a random sample. Although lift at 10% is most commonly used we have chosen to use lift at 0.5% and 1% since most of the datasets contain more than a million customers and 10% of customers is already too many for the company to approach in a retention campaign. The smaller portions will give a more realistic view of the performance.

Secondly, we will use the well known Area Under the Receiver Operating characteristics curve (AUC) which represents in a single number the trade off between specificity and sensitivity of the model [38].

Finally, although AUC is well known and widely used it has been shown to be incoherent when comparing different methods. Since the goal of this study is to compare methods, we opted to use the H-measure as well. The H-measure was introduced as a coherent alternative to AUC, since it compares all methods to the same metric [39]. In contrast to the AUC, the H-measure takes into account misclassification costs and in fact seeks to minimize them, to realize a value for performance.

IV. EXPERIMENTAL SET UP

For the experiments we have collected and analysed seven CDR datasets from around the world. The datasets together with some of their characteristics can be seen in table III. As the table shows, the number of customers in a network varies from a hundred thousand to over 4 million and the

TABLE III
DESCRIPTIONS OF DATASETS

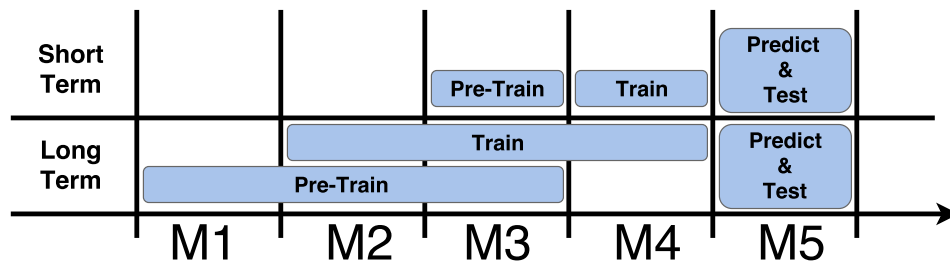| ID	Origin	Year	# Customers	Churn Rate	Contract Type
BC1	Belgium	2010	1.41 Million	4.4%	Prepaid
BC2	Belgium	2010	1.21 Million	0.84%	Postpaid
GD1	North America	2015	1.57 Million	0.71%	Postpaid
GD2	North America	2015	1.32 Million	2.5%	Prepaid
BP1	Europe	2008	4.33 Million	8.5%	Unknown
BP2	Europe	2008	4.52 Million	3.5%	Unknown
IS	Iceland	2015	93 Thousand	2.2%	Postpaid

Fig. 2. The figure shows how the datasets were split up to build the networks for pre-training, training and predicting in the short and long term setting.

churn rates are also quite varied, from under a percent to more than eight percent. The contract types are both prepaid and postpaid. Clearly, there is great variety in the datasets and their origin, which is beneficial when it comes to drawing general conclusions from our experiments.

All the datasets were preprocessed in the same way. Only users of the respective provider were considered and thus only within network correspondence. In addition, text messages were not taken into account, only phone calls and – based on an exploratory study – all phone calls lasting less than four seconds were disregarded, because they do not reflect the connection behaviour that we are modelling within the network.

As the CDR records for all of the datasets span six months, the experiment was set up to make the most of all that data. Based on previous studies [15], [4] and expert knowledge, churn was defined as being inactive for 30 days and then the churn day was defined as the day on which the customer became inactive. This also ensures the consistency between datasets, since actual churn dates were only available for a couple of them. Because of this definition, the last month can not be used for the analysis but only for building churn labels. This leaves five months of data to build the models and predict churn, which we name M1, M2, M3, M4 and M5, see Fig. 2.

We define two types of churn periods: long term and short term. The long term period is three months and the short term is one month. Two of the relational classifiers, CDRN and NLB, need to be pre-trained on a previous time frame to create reference vectors on which the classification is based. Therefore, a pre-train time frame was defined to have the same length as the train period but starting one month earlier. As a result, in the long term setting months M1, M2 and M3 are used for pre-training, months M2, M3 and M4 are used for training and in the short term setting, month M3 is used for pre training and month M4 is used for training. In both cases the prediction month is M5. See Fig. 2.

In addition, the networks were built using two types of edges, number of phone calls and length of phone calls. In both cases, the edges were weighted in time with decay as in equation 5.

Because of this setup, there were four different networks for each dataset to which each of the 24 relational learners were applied to. This resulted in 28 scores for each method. The scores from each of the relational learners were subsequently compared to the churn labels in month M5 to evaluate the models using the four performance measures, 0.5% lift, 1% lift, AUC and H-measure. Finally, various tests were applied to investigate the statistical significance of the results.

V. RESULTS

We will present the results to each of the research questions posed in section I. We follow the guidelines for statistical comparisons of classifiers over multiple datasets presented by [40]. To evaluate the statistical significance of the comparisons, we start by performing a non-parametric Friedman test, which tests the hypothesis that the average rank of all the methods is equal. Only when this hypothesis is rejected is it possible to continue with subsequent analyses to compare the individual methods. Here, we apply the post-hoc Nemenyi test [41] to compare the differences between all methods and to discover which differences are significant.

A. Comparison of Relational Learners

First we compare the performance of the 24 relational learners. To test for a significant difference between them, a Friedman test was applied to the rankings of each of the four performance measures. In all cases, the resulting p-value was close to zero, meaning that not all the methods are the same, and at least one of them performs significantly different from the rest.

Since the null hypothesis was rejected we continued by performing a post-hoc Nemenyi test to compare the differences between all methods. The results for the performance measures 0.5% lift, 1% lift, AUC and H-Measure can be seen in Fig. 3. In each of the subfigures, the line corresponding to each method represents its average rank. The higher the average rank, the further to the left this line is, with the left endpoint being the average rank itself and the right endpoint the critical difference from the average rank. The blue line represents the critical difference at the 95% confidence level of the best performing method. This means that if another line is situated fully to the right of the blue line, it performs significantly worse than the best performing method at the 95% confidence level. We see from Fig. 3 that for 0.5% lift twelve of the methods are significantly worse than the best one, fourteen for 1% lift, six for AUC and sixteen for the H-measure. Noticeably, the NLB performs better than the other relational classifiers and methods without a CI perform better as well.

B. Comparison of Relational Classifiers

In order to answer the second question regarding the best performing relational classifier, we apply a Friedman test. This
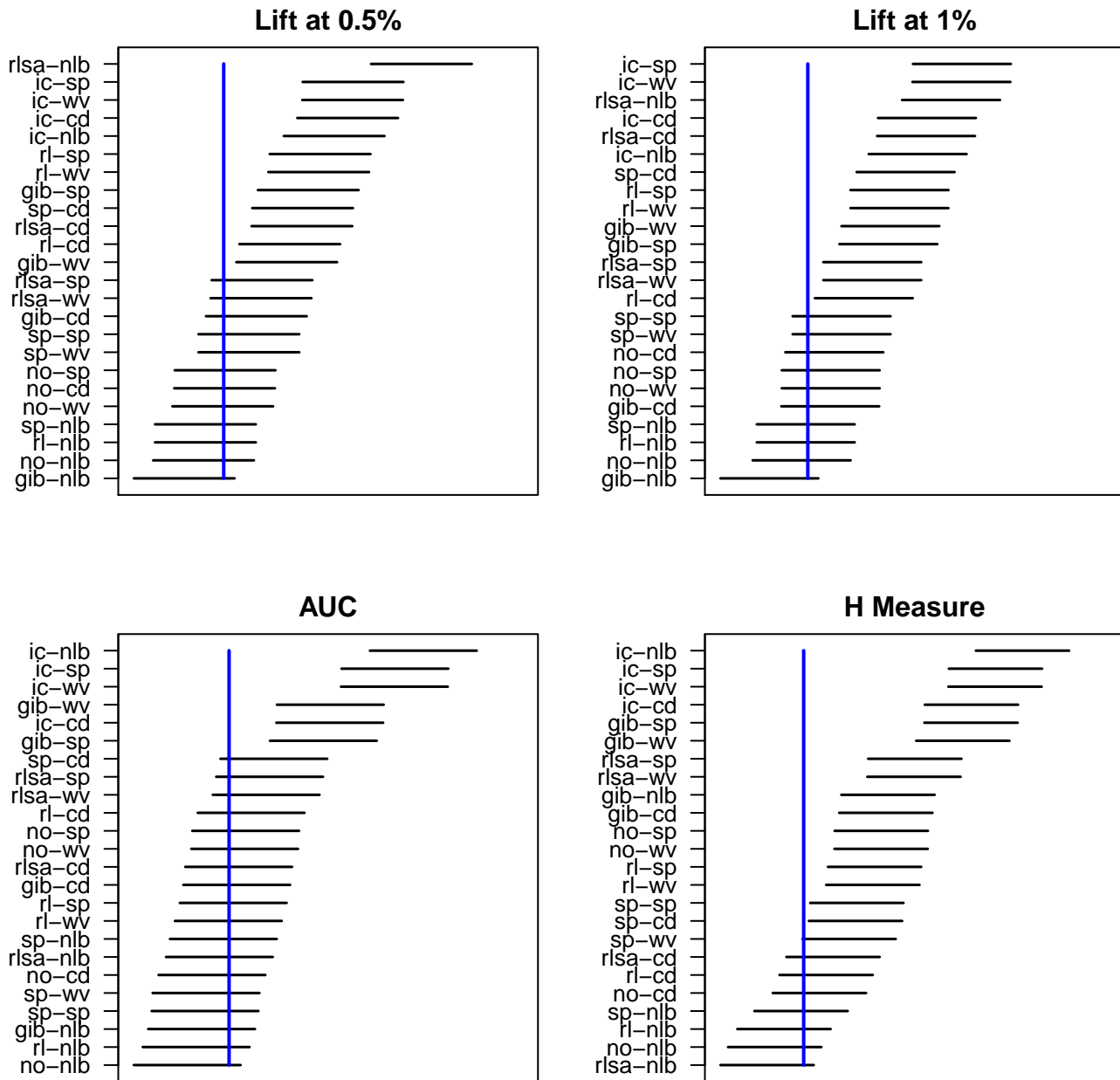
Fig. 3. The figures show the comparison of all methods against each other using the post-hoc Nemenyi test for the four performance measures. The best method is displayed in the bottom left corner of each figure with the methods that performed significantly worse on the right side of the vertical line which represents the 95% confidence level of the best performing method.

time, there are only four different methods, corresponding to the four classifiers. For all four measures the resulting p-value was less than 0.01, which means that there is some difference between the classifiers. Therefore, we proceed to compare them in a post-hoc Nemenyi test. The results can be seen in Fig. 4a which shows a boxplot of the differences in performance in 1% lift of each pair of the two classifiers. The grey boxes mean that the difference is significant at the 95% confidence level whereas the white colored boxes mean there is not a significant difference. As can be seen in the figure, the difference of NLB and any other classifier is always significant, which means that NLB performs better than the others classifiers.
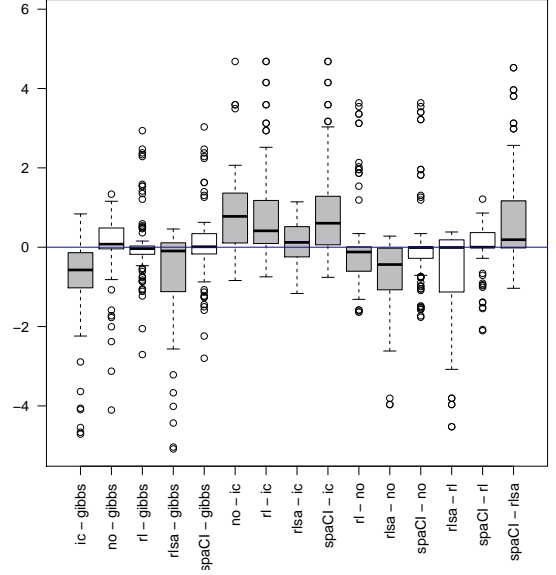
C. Comparison of Collective Inference Methods

The last step is to evaluate the differences of the collective inference methods. The Friedman test applied with the six collective inference methods gave a p-value of less than 0.001 for all of the four measures. The results of the subsequent post-hoc Nemenyi test on the differences, can be seen in Fig. 4b. It is evident from this figure that the IC method always performs significantly worse and not applying a collective inference method is often significantly better.

We conclude by investigating the impact of the collective inference methods in improving the performance of relational classifiers. The non-parametric Friedman test requires a balanced experimental design without repeated measurements and
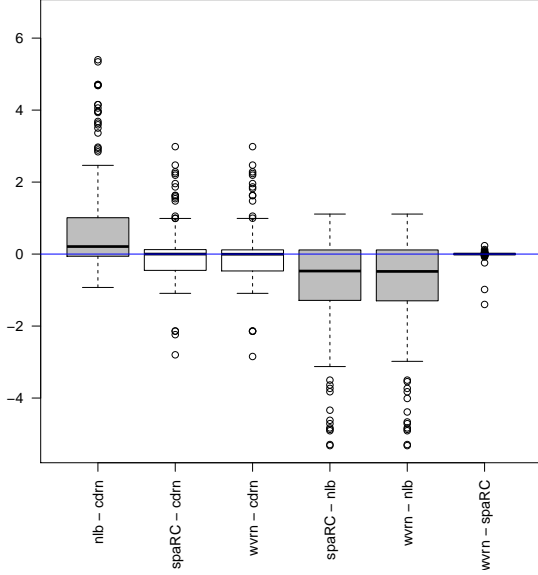
(a) Differences of Relational Classifiers

(b) Differences of Collective Inference Methods

Fig. 4. The figure shows boxplots of the differences in performance of the relational classifiers and the collective inference methods. Grey colored boxes represent statistically significant differences and the white boxes not statistically significant differences, at the 95% confidence level.

is therefore is not applicable in this case, since there are 20 methods with CI but only four without. Instead the non-parametric Kruskal-Wallis test is used. The null hypothesis, of the two samples originating from the same distribution, is rejected with a p-value of less than 0.01. The result is that methods without CI perform better than methods with CI.

VI. Conclusion

A. Main Findings

In this study, we have made a statistical comparison of the performance of two dozen relational learners when predicting churn in telco. In addition, we have compared the effects of the two components of the relational learners, namely relational classifiers and collective inference methods. Applying all these methods to 28 networks from seven datasets, we were able to receive robust results about the significance of the methods.

Firstly, we observed in the comparison of the relational learners that the same group of about a dozen learners consistently performs better than the rest. Our tests showed a clear separation between the better and the worse methods, measured by all applied performance measures.

Secondly, the statistical superiority of the network-only link-based classifier introduced by Lu and Getoor [21] was evident. This method builds a logistic regression model using link-based features extracted from the network. These features represent, for each node, the number of neighbors which have churned and not churned, whether a neighbor has churned and what the most common neighbor property is. As such, the classifier captures enough information from the network to accurately predict churn and this proves to be the best approach.

In addition, the iterative classification collective inference method, always performed worse than the other methods,

which might be due to the fact that after each step in the iteration labels (0 or 1) instead of scores (ranging form 0 to 1) are produced. As such, the result is much less refined than for the other CIs.

Finally, we have shown that in telco but possibly in other applications as well, collective inference methods do not improve the performance of relational classifiers when predicting churn. Explaining this is not straight forward and requires further research. One reason might be that since there are relatively few churn signals in the network, as well as edges, the signals get spread out and are not clear enough afterwards. Thus, the 'churn influence' gets too diluted to be meaningful.

B. Future Work

There is a whole range of possibilities for future research. So far, we have only used the scores from relational learners as predictions, but they could also be combined with local and network variables as attributes in models using classical binary classification techniques. The result would not only be better models, but it would also give a better idea of which learners are the most important. In addition, the results of relational and non-relational classifiers can be combined in different ways, which could be explored given the high number of datasets that we have.

Another aspect that could be investigated further is how people churn. Indeed, the way the networks for the analyses were built, offers the possibility of gaining a better understanding of the churn process itself. This could be a very interesting result for telco providers to better understand their clients.

Finally, we would like to mention the relatively novel maximum profit measure [38], [11]. It is specifically tailored to churn prediction problems and evaluates the models by taking into account the cost and benefits of a retention campaign. We

believe that applying the profit measure would give additional insight and important information when it comes to seeing how much the models are worth.

REFERENCES

[1] A. Berson and S. J. Smith, *Building data mining applications for CRM*. McGraw-Hill, Inc., 2002.

[2] W. Verbeke, D. Martens, C. Mues, and B. Baesens, "Building comprehensible customer churn prediction models with advanced rule induction techniques," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2354–2364, 2011.

[3] S. A. Macskassy and F. Provost, "Classification in networked data: A toolkit and a univariate case study," *The Journal of Machine Learning Research*, vol. 8, pp. 935–983, 2007.

[4] W. Verbeke, D. Martens, and B. Baesens, "Social network analysis for customer churn prediction," *Applied Soft Computing*, vol. 14, pp. 431–446, 2014.

[5] Y. Xie, X. Li, E. Ngai, and W. Ying, "Customer churn prediction using improved balanced random forests," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5445–5449, 2009.

[6] B. Larivière and D. Van den Poel, "Investigating the role of product features in preventing customer churn, by using survival analysis and choice modeling: The case of financial services," *Expert Systems with Applications*, vol. 27, no. 2, pp. 277–285, 2004.

[7] N. Glady, B. Baesens, and C. Croux, "Modeling churn using customer lifetime value," *European Journal of Operational Research*, vol. 197, no. 1, pp. 402–411, 2009.

[8] M. Guillén, J. P. Nielsen, T. H. Scheike, and A. M. Pérez-Marín, "Time-varying effects in the analysis of customer loyalty: A case study in insurance," *Expert Systems with Applications*, vol. 39, no. 3, pp. 3551–3558, 2012.

[9] C.-C. Günther, I. F. Tvete, K. Aas, G. I. Sandnes, and Ø. Borgan, "Modelling and predicting customer churn from an insurance company," *Scandinavian Actuarial Journal*, vol. 2014, no. 1, pp. 58–71, 2014.

[10] A. A. Khan, S. Jamwal, and M. Sepehri, "Applying data mining to customer churn prediction in an internet service provider," *International Journal of Computer Applications*, vol. 9, no. 7, pp. 8–14, 2010.

[11] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach," *European Journal of Operational Research*, vol. 218, no. 1, pp. 211–229, 2012.

[12] B. Baesens, *Analytics in a big data world: The essential guide to data science and its applications*. John Wiley & Sons, 2014.

[13] X. Zhang, J. Zhu, S. Xu, and Y. Wan, "Predicting customer churn through interpersonal influence," *Knowledge-Based Systems*, vol. 28, pp. 97–104, 2012.

[14] N. Modani, K. Dey, R. Gupta, and S. Godbole, "Cdr analysis based telco churn prediction and customer behavior insights: A case study," in *Web Information Systems Engineering–WISE 2013*. Springer, 2013, pp. 256–269.

[15] A. Backiel, B. Baesens, and G. Claeskens, "Mining telecommunication networks to enhance customer lifetime predictions," in *Artificial Intelligence and Soft Computing*. Springer, 2014, pp. 15–26.

[16] G. Benedek, Á. Lublóy, and G. Vastag, "The importance of social embeddedness: Churn models at mobile providers," *Decision Sciences*, vol. 45, no. 1, pp. 175–201, 2014.

[17] K. Dasgupta, R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjea, A. A. Nanavati, and A. Joshi, "Social ties and their relevance to churn in mobile telecom networks," in *Proceedings of the 11th international conference on Extending database technology: Advances in database technology*. ACM, 2008, pp. 668–677.

[18] A. Backiel, Y. Verbinnen, B. Baesens, and G. Claeskens, "Combining local and social network classifiers to improve churn prediction," in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM, 2015, pp. 651–658.

[19] P. D. Kusuma, D. Radosavljevik, F. W. Takes, and P. van der Putten, "Combining customer attribute and social network mining for prepaid mobile churn prediction," in *Proc. the 23rd Annual Belgian Dutch Conference on Machine Learning (BENELEARN)*, 2013, pp. 50–58.

[20] K. Kim, C.-H. Jun, and J. Lee, "Improved churn prediction in telecommunication industry by analyzing a large network," *Expert Systems with Applications*, vol. 41, no. 15, pp. 6575–6584, 2014.

[21] Q. Lu and L. Getoor, "Link-based classification," in *ICML*, vol. 3, 2003, pp. 496–503.

[22] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual review of sociology*, pp. 415–444, 2001.

[23] L. Kovanen, M. Karsai, K. Kaski, J. Kertész, and J. Saramäki, "Temporal motifs in time-dependent networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2011, no. 11, p. P11005, 2011.

[24] C. Sarraute, P. Blanc, and J. Burroni, "A study of age and gender seen through mobile phone usage patterns in mexico," in *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*. IEEE, 2014, pp. 836–843.

[25] C. Rhodes and P. Jones, "Inferring missing links in partially observed social networks," *Journal of the operational research society*, vol. 60, no. 10, pp. 1373–1383, 2009.

[26] C. R. Shalizi and A. C. Thomas, "Homophily and contagion are generically confounded in observational social network studies," *Sociological methods & research*, vol. 40, no. 2, pp. 211–239, 2011.

[27] H.-Y. Liao, K.-Y. Chen, D.-R. Liu, and Y.-L. Chiu, "Customer churn prediction in virtual worlds," in *Advanced Applied Informatics (IIAI-AAI), 2015 IIAI 4th International Congress on*. IEEE, 2015, pp. 115–120.

[28] V. Van Vlasselaer, T. Eliassi-Rad, L. Akoglu, M. Snoeck, and B. Baesens, "Gotcha! network-based fraud detection for social security fraud," *under peer review*, 2014.

[29] V. Van Vlasselaer, C. Bravo, O. Caelen, T. Eliassi-Rad, L. Akoglu, M. Snoeck, and B. Baesens, "Apate: A novel approach for automated credit card transaction fraud detection using network-based extensions," *Decision Support Systems*, vol. 75, pp. 38–48, 2015.

[30] M. Newman, *Networks: an introduction*. OUP Oxford, 2010.

[31] J. J. Rocchio, "Relevance feedback in information retrieval," 1971.

[32] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad, "Collective classification in network data," *AI magazine*, vol. 29, no. 3, p. 93, 2008.

[33] D. Jensen, J. Neville, and B. Gallagher, "Why collective inference improves relational classification," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 593–598.

[34] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 6, pp. 721–741, 1984.

[35] S. Chakrabarti, B. Dom, and P. Indyk, "Enhanced hypertext categorization using hyperlinks," in *ACM SIGMOD Record*, vol. 27, no. 2. ACM, 1998, pp. 307–318.

[36] A. N. Langville and C. D. Meyer, *Google's PageRank and beyond: The science of search engine rankings*. Princeton University Press, 2011.

[37] S.-Y. Hung, D. C. Yen, and H.-Y. Wang, "Applying data mining to telecom churn management," *Expert Systems with Applications*, vol. 31, no. 3, pp. 515–524, 2006.

[38] T. Verbraken, W. Verbeke, and B. Baesens, "A novel profit maximizing metric for measuring classification performance of customer churn prediction models," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 25, no. 5, pp. 961–973, 2013.

[39] D. J. Hand, "Measuring classifier performance: a coherent alternative to the area under the roc curve," *Machine learning*, vol. 77, no. 1, pp. 103–123, 2009.

[40] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[41] P. Nemenyi, "Distribution-free multiple comparisons," in *Biometrics*, vol. 18, no. 2, 1962, p. 263.