Churn prediction using representation learning with guided random walks

Sandra Mitrović[1], Jochen De Weerdt[1],
[1] Department of Decision Sciences and Information Management, KU Leuven, Belgium
sandra.mitrovic@kuleuven.be, jochen.deweerdt@kuleuven.be

June 16, 2018

Abstract

Unleashing the full potential of data is oftentimes a cumbersome task, especially when dealing with network data. It is therefore possible that while focusing on one part of the solution, other valuable pieces of information remain under-treated leading to under-performing results. In this work, we zoom into the nature of an augmentation of call graphs devised for addressing churn prediction in telco. By shifting the focus from a homogeneous to a heterogeneous perspective, by defining different probabilistic meta paths, and by applying representation learning on graphs using these defined meta paths, we demonstrate the benefits of this approach, not only by means of improvements of predictive results, but also with promising insights regarding the interplay of meta path type and predictive outcome.

1 Introduction

Predictive analytics suffers from different problems regarding data: the lack of (appropriate) data, bad quality of data etc. Not less important is the inability to leverage the available data and provide more structural ways of handling each phase of the predictive process. As for the other predictive tasks, this also holds for predicting which customers will stop using telecommunication operator services, known in the literature as *telco churn prediction*. Actually, the problem gets even more emphasized in this setting, given that Call Detail Record (CDR) data can be transformed into networks/graphs[1], which are quite cumbersome to handle due to their complex structure and large size. Therefore, related studies typically encounter problems of under-exploiting available data and/or hand-crafting variables as e.g. in [Chen *et al.*, 2012; Kim *et al.*, 2014]. A recently proposed study [Mitrović *et al.*, 2017] suggested an approach to deal with these problems by first, augmenting CDR graphs in order to join all available information and thus enforce unleashing the full potential of the data and second, applying representation learning on such graphs in order to avoid hand-crafting. More precisely, the aforementioned augmented CDR graphs are constructed by deriving artificial nodes based on the RFM (**R**ecency, **F**requency, **M**onetary) model [Hughes, 1994] and connecting them to customers present in the original CDR graph, as appropriate. Next, a representation learning approach, similar to other methods like DeepWalk [Perozzi *et al.*, 2014] and Node2vec [Grover and Leskovec, 2016], is applied. It attempts to maximize the probability that nearby nodes (or in other words: nodes having the same contexts, where context is defined by a random walk) will have similar representations. Obviously, constructed augmented graphs contain different types of nodes and as such, fit into the definition of heterogeneous graphs [Sun *et al.*, 2011], even though they are not treated in that way in [Mitrović *et al.*, 2017]. Since different node types convey diverse information and given the applied representation learning method, guiding random walks in order to generate more specifically tailored contexts to account for (and compare) these differences, seems worthwhile. Furthermore, previous studies have shown that taking into account heterogeneous information can improve predictive accuracy for classification (and clustering) tasks [Sun *et al.*, 2011; Dong *et al.*, 2017; Fu *et al.*, 2017].

Therefore, in this work we exploit the same kind of networks while adjusting the representation learning part to accommodate for heterogeneous networks, using meta paths, similarly as has been done in [Dong *et al.*, 2017]. However, unlike the meta paths defined in [Dong *et al.*, 2017], we allow for probabilistic and asymmetrical meta paths.

Our results show that representation learning on heterogeneous graphs using different meta paths can be beneficial for churn prediction as compared to representation learning on homogeneous graphs. Moreover, using different types of meta paths leads to interesting insights which provide motivation for further exploration of the interplay between different interactions, corresponding meta paths and the predictive result.

The rest of the paper is organized as follows: in Sections 2 and 3, we provide a short overview of related literature and preliminaries, respectively; in Section 4 we elaborate our methodology, while in Section 5 we explain our experimental setup; Section 6 provides the results of our experiments and finally, Section 7 concludes our study with several directions for future research.

2 Related Work

Churn prediction in telco has so far been addressed in the literature from different perspectives, using different feature types and models. In this short overview we limit ourselves mainly to representation learning-based methods. These

[1] taking customers as nodes and their interactions as edges

methods have initially achieved great success for word embedding in the natural language processing (NLP) domain [Bengio *et al.*, 2003; 2013; Mikolov *et al.*, 2013] but were recently transferred to graphs primarily by [Perozzi *et al.*, 2014] but also [Tang *et al.*, 2015b; Grover and Leskovec, 2016; Cao *et al.*, 2016; Dong *et al.*, 2017]. For the purpose of predicting churn in telco, explicit learning of customer representations was very rarely used. However, deep learning methods were used in several studies [Castanedo *et al.*, 2014; Wangperawong *et al.*, 2016; Umayaparvathi and Iyakutti, 2017] although not with too much success apart from [Castanedo *et al.*, 2014] which however does not reveal all the details of dataset used.

The idea of this work is to take advantage of the synergy of two existing works, i.e. an approach to apply explicit customer representation learning to churn prediction in [Mitrović *et al.*, 2017], and the work of [Dong *et al.*, 2017], which proposes to use meta paths for representation learning in heterogeneous graphs. In [Mitrović *et al.*, 2017], an adaptation of node2vec [Grover and Leskovec, 2016] was developed to make it scalable on large graphs obtained from CDR data. Node2vec itself extends on DeepWalk [Perozzi *et al.*, 2014] by defining more flexible random walks, which, however, come with additional computational burden. Furthermore, [Mitrović *et al.*, 2017] overcomes several problems known in the churn prediction domain, such as feature hand-crafting and under-exploiting of available data, as is the case in e.g. [Chen *et al.*, 2012; Kim *et al.*, 2014]. Given that augmented graph architectures proposed in [Mitrović *et al.*, 2017] enable integrating both information related to customer interaction as well as information conveyed by pure graph topology, we resort to using the same graph constructs. However, the drawback of the approach in [Mitrović *et al.*, 2017] is in applying representation learning on RFM-Augmented graphs treating them as homogeneous. While the idea of applying representation learning on RFM-Augmented graphs remains valid, given that RFM-Augmented graphs are inherently heterogeneous in nature, the option of using a meta path based representation learning method as presented in [Dong *et al.*, 2017] seems attractive. The latter method enables guiding random walks and is more appropriate for heterogeneous graphs. Still, we slightly adapt it to our setting by defining probabilistic version of random walks and not requiring walk symmetry as in [Dong *et al.*, 2017].

Other works on heterogeneous networks include [Sun *et al.*, 2011; Chang *et al.*, 2015; Tang *et al.*, 2015a; Chen and Sun, 2017; Fu *et al.*, 2017]. Work of [Sun *et al.*, 2011] is also based on meta paths but it focuses on defining a similarity measure instead of learning representations. In [Chang *et al.*, 2015] learning representations is done by deep convolutional neural networks, while in [Tang *et al.*, 2015a] the focus is on text embeddings. Similar to our work, [Chen and Sun, 2017; Fu *et al.*, 2017] leverage particular meta paths in heterogeneous networks based on the predictive goal. However, the particularities of both methods differ from our work. Moreover, in [Chen and Sun, 2017] the goal is author identification while in [Fu *et al.*, 2017] learning is performed not only on the node, but also on the link level, which is not of interest for this work.

3 Preliminaries

In this section, we will first provide several definitions which will be useful for following further elucidation of our methodology. Next, we provide a brief explanation of the construction of call networks from CDRs, as done in [Mitrović *et al.*, 2017].

3.1 Basic Definitions

Definition 3.1. *(RFM Model)* The RFM Model as proposed by [Hughes, 1994] explains customer behavior in an observed time period with respect to a certain event based on the following three variables:

- **R**ecency, which quantifies how recent the last event was of a customer during the observed time interval;
- **F**requency, which quantifies how frequently a customer experienced the particular event during the observed time interval;
- **M**onetary, which quantifies the monetary amount a customer spent during the observed time interval in relation to this particular event.

In our context, the particular event is a phone call but many other interpretations such as a purchase, a credit card transaction, etc. are possible, depending on the particular domain.

Definition 3.2. *(Heterogeneous network)* A graph $G = (V, E)$ is called a heterogeneous graph if there exists a node type mapping function $\phi : V \rightarrow T_V$ and an edge type mapping function $\psi : E \rightarrow T_E$ such that either $|T_V| > 1$ or $|T_E| > 1$. We will use notation $G = (V, E, T_V, T_E, \phi, \psi)$ for such a graph.

Definition 3.3. *(Probabilistic meta path)* A path class $\mathcal{P}$ constructed over a heterogeneous graph $G = (V, E, T_V, T_E, \phi, \psi)$ which allows graph traversal through different node and edge types and can be written in the form $T_{v_1} \xrightarrow{T_{e_1}} ... \xrightarrow{T_{e_{i-1}}} T_{v_i} \xrightarrow{T_{e_i}} T_{v_j} \xrightarrow{T_{e_j}} ... \xrightarrow{T_{e_{k-1}}} T_{v_k}$ where $T_{v_i}, T_{v_j} \in T_V$, $T_{e_i}, T_{e_j} \in T_E$, with the transition probability at step i defined as:

$$Prob(T_{v_j}|T_{v_i}, \mathcal{P}) = \begin{cases} p_{reg}, & \text{if } T_{v_j} \neq T_{v_i} \\ p_{min}, & \text{otherwise} \end{cases}$$

where $p_{min} \ll p_{reg}$, is called a probabilistic meta path. In other words, we define the types of nodes which should be visited along the path with certain (high) probability p_{reg}, but with a very small probability p_{min} we allow diverging from the set path.

Definition 3.4. *(Probabilistic meta path instance)* A path $P \in \mathcal{P}$ of the form $v_1 \xrightarrow{e_1} ... \xrightarrow{e_{i-1}} v_i \xrightarrow{e_i} v_j \xrightarrow{e_j} ... \xrightarrow{e_{k-1}} v_k$ constructed over a heterogeneous graph $G = (V, E, T_V, T_E, \phi, \psi)$, where $v_i \in V$, $e_i \in E$, $\phi(v_i) = T_{v_i}$, $\phi(v_j) = T_{v_j}$, $\psi(e_i) = T_{e_i}$, $\phi(e_j) = T_{e_j}$, $T_{v_i} \neq T_{v_j}$ and $T_{e_i} \neq T_{e_j}$ is called an instance of probabilistic meta path $\mathcal{P}$ (a path following probabilistic meta path $\mathcal{P}$). In cases when the edge type is not crucial, a notation $P = (v_1, ..., v_i, v_j, ..., v_k)$ will be used.

Previous works (as e.g. [Dong *et al.*, 2017]) define meta paths[2] to be symmetric ($T_{v_1} = T_{v_k}$) and without the possibility of taking alternate paths (hence $p_{min} = 0$). Moreover, to accommodate for particularities of the data and problem at hand, we will allow for probabilistic meta paths where T_{v_j} is not necessarily different from T_{v_i} (but the alternative option would still be possible only with a very small probability). The main motivation for introducing probabilistic meta paths is to ensure existence of meta path instances of sufficient length (as this might be a problem due to graph sparsity). In the rest of the paper, we always refer to a probabilistic meta path even if we often use meta path for better readability.

3.2 Constructing RFM-Augmented Networks from Telco CDRs

In [Mitrović *et al.*, 2017], a special type of call graph architectures, referred to as RFM-Augmented graphs have been constructed and their benefits for churn prediction in telco have been proven as compared to traditional RFM variables. We therefore opt for the same call graph architecture and provide a brief explanation of these.

The main idea of RFM-Augmented graphs is to augment the original call graph with artificial nodes which are generated based on RFM variables. Hence, it is important to define, first, how these RFM variables are calculated and second, how artificial nodes are generated based on those variables.

The Choice of RFM Variables

Despite the presence of a variety of RFM variables in the literature, to this end, only three different versions of RFM variables are considered: 1) Summary-RFM (denoted by RFM_s) whereby RFM variables are calculated considering all customer interactions; 2) Detailed-RFM (denoted by RFM_d), whereby customer interactions are divided according to call direction and destination into three subcategories: a) outgoing towards home network, b) outgoing towards other networks and c) incoming and each of R, F, M is calculated according to these; and 3) Churn-RFM (denoted by RFM_{ch}) whereby RFM variables are calculated only with respect to interactions with customers who are known to have churned. The latter is motivated by the fact that previous works on churn prediction in telco have recognized the importance of connections with already churned customers [Dasgupta *et al.*, 2008; Modani *et al.*, 2013; Zhang *et al.*, 2012].

Generation of Artificial Nodes

Once the R/F/M variables are derived for each node, each of these R/F/M variables is partitioned in five groups (corresponding to very high, high, medium, low, very low) similar to procedures already used in the literature [Hughes, 1994; McCarty and Hastak, 2007; Cheng and Chen, 2009]. Each of these groups is then assigned one artificial node. In case of Summary-RFM (RFM_s), given that each node is assigned one variable per R, F, M, described partitioning results in 15 artificial nodes R_i, F_i, M_i, where $i \in \{1, 2, ..5\}$. Next, nodes from the original call graph are connected to artificial nodes according to their corresponding R/F/M partitions. Such an architecture is denoted by AG_s. A similar set of steps is used to construct the second type of architecture, except that the procedure does not start with Summary-RFM variables, i.e. one variable per R, F, M (per node), but instead with Detailed-RFM (RFM_d) hence, three different variables per each of R, F, M (per node) which are further on partitioned (each of them separately) into five groups. This graph construction is denoted by AG_d. An additional augmentation of both of the previous two graph types is obtained by adding artificial churn node to which all identified churners are then connected. The obtained RFM-Augmented graphs are denoted by AG_{s+ch} and AG_{d+ch} for AG_s and AG_d graph, respectively.

[2]Previous works also use notion of relation instead of edge type which turns a meta path into a composite relation of different relation types [Sun and Han, 2013; Dong *et al.*, 2017].

4 Methodology

In this section, we will explain the construction of different types of guided random walks, and elaborate on representation learning procedure.

4.1 Defining Meta Paths in RFM-Augmented Networks

In [Mitrović *et al.*, 2017], the authors draw attention to different types of nodes and links present in RFM-Augmented graphs and use that as motivation for considering these graphs as unweighted. However, in the proceedings, the representation learning method based on random walks does not make a distinction between different node types. Hence, essentially their method treats RFM-Augmented graphs as homogeneous. Nevertheless, by looking into Definition 3.2 we can see that these graphs are inherently heterogeneous, with $|T_V| = |T_E| = 2$ for AG_s and AG_d and $|T_V| = |T_E| = 3$ for AG_{s+ch} and AG_{d+ch} (for AG_s and AG_d, we can perceive that the nodes belonging to the original topology are connected by edges of one type, while the same nodes are connected to the artificial ones with another type of edges).

We will use the common notation $AG_* = (V, E, T_V, T_E, \phi, \psi)$ for all RFM-Augmented graphs, where $* \in \{s, s+ch, d, d+ch\}$ and $T_V = \{CN \cup AN\}$ with CN denoting the set of customer nodes and AN denoting the set of all artificial nodes. Additionally, in graphs AG_{s+ch} and AG_{d+ch}, we will make a distinction between a set of artificial RFM nodes, denoted as AN_{rfm}, and a set containing the single artificial churn node, denoted as AN_{ch}. Hence, $AN = AN_{rfm} \cup AN_{ch}$. As paying special attention to edge types (T_E) will not be necessary, we will further shorten the notation to $AG_* = (V, E, \{CN \cup AN\}, \phi)$ where $* \in \{s, d\}$ and $AG_* = (V, E, \{CN \cup AN_{rfm} \cup AN_{ch}\}, \phi)$, for $* \in \{s+ch, d+ch\}$.

In heterogeneous networks, on the contrary to homogeneous ones, nodes can be connected via different types of paths, known as meta paths defined in Definition 3.3. Moreover, different node and edge types convey diverse information, and taking that explicitly into account can enhance predictive performance, as shown in previous works, e.g. [Sun *et al.*, 2011]. Therefore, in order to leverage RFM-Augmented

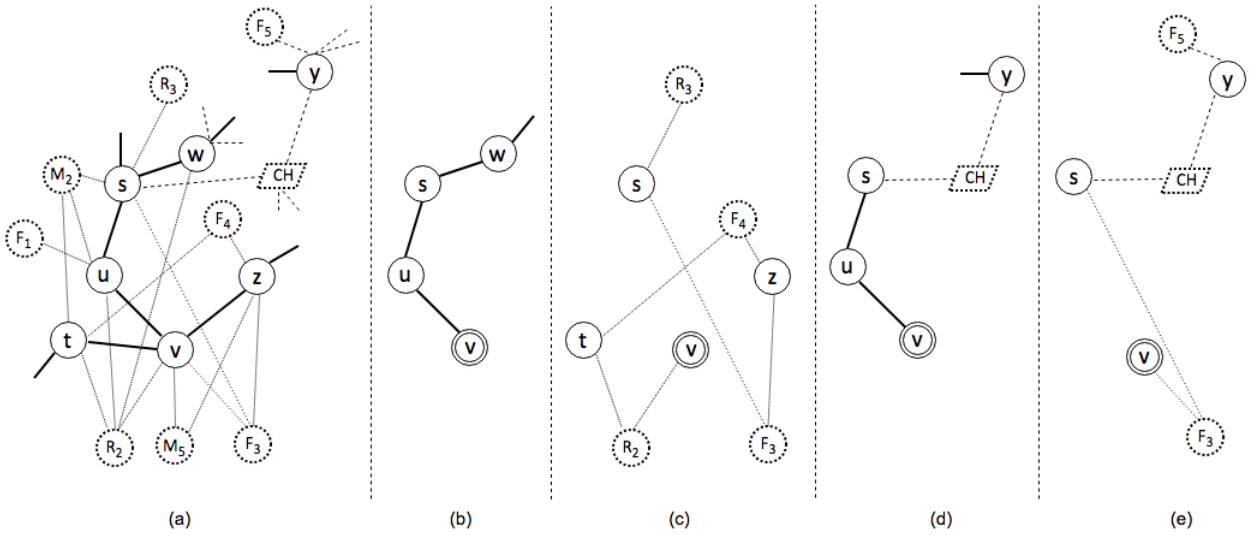

Figure 1: Graphical illustration of defined meta paths on the AG_{s+ch} RFM-Augmented network. In a) a fragment of the network depicting the neighborhood of node v. Nodes of type CN are represented by solid line-edged circles. Nodes of type AN_{rfm} are denoted by circles with dotted edges.The artificial churn node (of type AN_{ch}) is a romb with dotted edges. Connections with artificial nodes (AN) are dotted while edges interconnecting nodes of CN type are solid lines. In b)-e) instances of defined meta paths are represented (all starting from node v). Particularly: in b) an example of a local meta path; in c) an instance of an alternating meta path type; in d) instance of a local-churn meta path and finally, in e) an instance of an alternating-churn meta path.

graphs, it is reasonable to define different types of probabilistic meta paths, as follows:

- A meta path $\mathcal{P}_{\mathcal{C}}$ constructed over an RFM-Augmented graph $AG_* = (V, E, \{CN \cup AN\}, \phi)$ of the form $CN \rightarrow CN \rightarrow CN \rightarrow ....$ Hence, this meta path discriminates in favor of the CN node type at each step (to be referred to as *local* meta path). In other words, we consider meta path instances $P_C \in \mathcal{P}_{\mathcal{C}}$ of the form $P_C = (v_1, ..., v_i, v_{i+1}, ..., v_k)$, $(v_i, v_{i+1}) = e_i \in E$ where transition probability at step i is defined as:

$$Prob(v_{i+1}|v_i, \mathcal{P}_{\mathcal{C}}) = \begin{cases} 0.99, & \text{if } \phi(v_{i+1}) = CN \\ 0.01, & \text{if } \phi(v_{i+1}) = AN \end{cases}$$

- A meta path $\mathcal{P}_{\mathcal{A}}$ constructed over an RFM-Augmented graph $AG_* = (V, E, \{CN \cup AN\}, \phi)$ of any of the two forms: $CN \rightarrow AN \rightarrow CN \rightarrow$ ... or $AN \rightarrow CN \rightarrow AN \rightarrow ....$ Hence, this meta path favors alternation between CN and AN node types at each step (to be referred to as *alternating* meta path). In other words, we consider meta path instances $P_A \in \mathcal{P}_{\mathcal{A}}$ of the form $P_A = (v_1, ..., v_i, v_{i+1}, ..., v_k)$, $(v_i, v_{i+1}) = e_i \in E$ where transition probability at step i is defined as:

$$Prob(v_{i+1}|v_i, \mathcal{P}_{\mathcal{A}}) = \begin{cases} 0.99, & \text{if } \phi(v_i) \neq \phi(v_{i+1}) \\ 0.01, & \text{otherwise} \end{cases}$$

- A meta path $\mathcal{P}_{\bar{\mathcal{C}}}$ constructed over an RFM-Augmented graph $AG_* = (V, E, \{CN \cup AN_{rfm} \cup AN_{ch}\}, \phi)$ of the form $CN \rightarrow AN_{ch} \rightarrow CN \rightarrow ....$ Hence, this meta path discriminates in favor of the artificial churn node AN_{ch} after each CN node type, whenever possible, that is, if such link exists. This meta path will be referred to as *local-churn* meta path. In other words, we consider meta path instances $P_{\bar{C}} \in \mathcal{P}_{\bar{\mathcal{C}}}$ of the form $P_{\bar{C}} = (v_1, ..., v_i, v_{i+1}, ..., v_k)$, $(v_i, v_{i+1}) = e_i \in E$ where transition probability at step i is defined as:

$$Prob(v_{i+1}|v_i, \mathcal{P}_{\bar{\mathcal{C}}}) = \begin{cases} 0.90, & \text{if } \phi(v_{i+1}) = AN_{ch} \\ 0.09, & \text{if } \phi(v_{i+1}) = CN \\ 0.01, & \text{if } \phi(v_{i+1}) = AN_{rfm} \end{cases}$$

- A meta path $\mathcal{P}_{\bar{\mathcal{A}}}$ constructed over an RFM-Augmented graph $AG_* = (V, E, \{CN \cup AN_{rfm} \cup AN_{ch}\}, \phi)$ of any of the two forms: $CN \rightarrow AN_{ch} \rightarrow CN \rightarrow AN_{rfm} \rightarrow CN \rightarrow$ ... or $AN_{rfm} \rightarrow CN \rightarrow AN_{ch} \rightarrow CN \rightarrow AN_{rfm} \rightarrow CN \rightarrow ....$ Hence, this meta path favors visiting churn node type AN_{ch} while alternating between CN and AN node types, whenever possible, that is, if such link exists. This meta path will be referred to as *alternating-churn* meta path. In other words, we consider meta path instances $P_{\bar{A}} \in \mathcal{P}_{\bar{\mathcal{A}}}$ of the form $P_{\bar{A}} = (v_1, ..., v_i, v_{i+1}, ..., v_k)$, $(v_i, v_{i+1}) = e_i \in E$ where transition probability at step i is defined as:

$$Prob(v_{i+1}|v_i, \mathcal{P}_{\bar{\mathcal{A}}}) = \begin{cases} 0.90, & \text{if } \phi(v_{i+1}) = AN_{ch} \\ 0.09, & \text{if } \phi(v_{i+1}) \neq AN_{ch} \\ & \text{and } \phi(v_i) \neq \phi(v_{i+1}) \\ 0.01, & \text{if } \phi(v_{i+1}) \neq AN_{ch} \\ & \text{and } \phi(v_i) = \phi(v_{i+1}) \end{cases}$$

Obviously, *-churn meta paths are not to be considered on AG_s and AG_d RFM-Augmented graphs.

Figure 1 depicts one fragment of an RFM-Augmented AG_{s+ch} network (in a)) along with path instances for local

meta path (in b)), alternating meta path (in c)), local-churn meta path (in d)) and alternating-churn meta path (in e)).

4.2 Meta path based Representation Learning on RFM-Augmented Networks

As already explained in Section 2, our approach is similar to the metapath2vec approach in [Dong *et al.*, 2017], which in turn is based on SkipGram model from NLP domain and the ideas of its adaptation on graph setting from [Perozzi *et al.*, 2014; Grover and Leskovec, 2016].

SkipGram itself is a maximum likelikood optimization problem of finding a low (d-)dimensional representation f for each word v of vocabulary V such that the probability of predicting its nearby words (within its corresponding context C_v) is maximized. That is, the model attempts to find $f, f : V \rightarrow R^d, d \ll |V|$ such that

$$max \sum_{v \in V} \sum_{w \in C_v} logPr(w|f(v)).$$

The background hypothesis of SkipGram states that words are more similar the more they appear in similar contexts (known as distributional hypothesis). Given that similar hypothesis exists in the graph setting in the form of homophily, the SkipGram idea can be transferred to graphs, by redefining the concept of word context into a concept of node neighbourhood. Node neighborhood is usually defined by means of truncated random walks starting from the node itself. Moreover, in heterogeneous networks, we can use the concept of meta paths to define different types of context. Then the above objective becomes:

$$\max_f \sum_{v \in V} \sum_{w_P \in C_v^P} logPr(w_P|f(v)),$$

where C_v^P is a context of a node v defined by a meta path P. The conditional probability $Pr(w|f(v))$ usually defined by a soft-max function becomes $Pr(w_P|f(v)) = \frac{\exp(f(w_P)\cdot f(v))}{\sum_{u \in V} \exp(f(u)\cdot f(v))}$. Given the complexity of the denominator and the goal (which is essentially to find useful representations and not to find the exact optimum of the function), negative sampling is used to approximate the objective, thus leading to a new objective being

$$\max_f \sum_{w_P \in C_v^P} log\ \sigma(-f(w_P)\cdot f(v)) + \sum_{n_P \notin C_v^P} log\ \sigma(-f(n_P)\cdot f(v)),$$

where σ is the sigmoid function.

Node representations obtained by the previously explained procedure are used as input features for an $l2-$regularized logistic regression classifier aiming to predict churn.

5 Experimental setup

We use the same two datasets (one prepaid, one postpaid) as in [Mitrović *et al.*, 2017], hence we refer an interested reader to that work for the datasets' details. Both datasets are only CDR-based which on one hand limits the diversity of available information, but on the other hand ensures the possibility for reproducing our method on other datasets (as any telco operator would have the information used here at their disposal, in contrast to some other churn prediction approaches which tend to avoid disclosure of the complete information used, like [Castanedo *et al.*, 2014], or opt for using not so easily available data e.g. customer complaints in [Huang *et al.*, 2015]).

Additionally we use the results obtained in [Mitrović *et al.*, 2017] as our baselines, and therefore, to obtain a fair comparison, we mimic the same experimental setup and churn definition. This includes setting the same parameters for random walks (walk length 30, number of walks 10) and SkipGram model (number of dimensions 128, window length 10, number of negative samples 5) as in [Mitrović *et al.*, 2017].

For the baselines we choose both traditional RFM variables as well as the results obtained in [Mitrović *et al.*, 2017] on all four RFM-Augmented networks $AG_s, AG_d, AG_{s+ch}, AG_{d+ch}$. We compare these with four different scenarios, based on four differently defined meta paths.

6 Results

Results obtained using different meta paths on both datasets and different kinds of RFM-Augmented networks can be seen in Table 1. It is obvious that an alternating meta path-based approach always outperforms a local-based approach, in terms of AUC. Likewise, in terms of AUC, the alternating-churn meta path-based approach performs better than the local-churn one. Our local meta path definition reflects homophily[3] based on customer interactions. Given that alternating meta path reflects the factual similarity of customers measured by R, F, M variables, we can infer that homophily indeed plays an important role for predicting churn, which confirms findings of previous studies [Verbeke *et al.*, 2014].

Additionally, we can notice that for AG_s, in terms of AUC, the learning approach based on alternating meta paths outperforms all the other methods. On the contrary, in case of lift, the baseline methods still perform better.

Finally, we would like to discuss the effect that adding churn information has on the obtained AUC scores. Our previous study indicates that switching from AG_s to AG_{s+ch} for random walk generation yields better AUC scores (see baseline figures). On the one hand, our results based on both local and alternating heterogeneous meta paths, confirm this finding. On the other hand, if churn information indeed increases performance, we would expect that local-churn and alternating-churn meta path-based AUC scores outperform local and alternating ones, respectively. As this, however, does not seem to be the case, we performed a short comparison of generated local and local-churn meta path-based walks for prepaid AG_{s+ch}. The total number of generated walks in both cases is the same (43035570), however, the presence of AN_{ch} and AN_{rfm} nodes is remarkably different: with respect to total number of walks, AN_{rfm} nodes occur in 25.16% in case of local and 24.20% in case of local-churn meta path-based walks, while AN_{ch} occur in only 0.15% in case of local and 19.99% in case of local-churn meta path-based walks.

[3]Homophily is a tendency to link to those who are perceived as similar.

Table 1: Comparison in terms of AUC and lift (at 0.5%, between parenthesis) among different methods for prepaid (upper) and postpaid (lower) datasets. The results are averaged across 10 folds (different from folds used for hyperparameter tuning). Baselines are marked with the star symbol. The best AUC score per RFM-Augmented graph type per dataset is marked in bold. The best overall AUC score per dataset is marked in bold and underlined.

Dataset	Graph Type	RFM Variables*	Homogeneous*	Heterogeneous: meta path-based			
				Local	Alternating	Local-churn	Alternating-churn
Prepaid	AG_s	0.668 (1.993)	0.679 (2.070)	0.6383 (1.8688)	**0.6799** (1.8943)	n/a	n/a
	AG_d	**0.687** (2.087)	0.666 (1.953)	0.6286 (1.7963)	0.6729 (1.8570)	n/a	n/a
	AG_{s+ch}	0.669 (1.995)	**0.698** (2.439)	0.6384 (1.8369)	0.6857 (2.0243)	0.6341 (1.9213)	0.6840 (1.9130)
	AG_{d+ch}	0.686 (2.087)	**0.699** (2.500)	0.6297 (1.7968)	0.6762 (1.8872)	0.6260 (1.8557)	0.6780 (1.9062)
Postpaid	AG_s	0.726 (3.784)	0.748 (4.153)	0.6254 (2.7350)	**0.7525** (3.6775)	n/a	n/a
	AG_d	**0.744** (4.322)	0.732 (3.792)	0.6107 (2.5211)	0.7405 (3.7592)	n/a	n/a
	AG_{s+ch}	0.727 (3.779)	0.748 (4.430)	0.6263 (2.7504)	**0.7526** (3.8083)	0.6215 (2.6957)	0.7514 (3.8182)
	AG_{d+ch}	**0.744** (4.322)	0.733 (3.922)	0.6101 (2.5076)	0.7408 (3.7710)	0.6118 (2.5043)	0.7403 (3.8598)

Even though it might look counter-intuitive, it seems that the presence of AN_{rfm} contributes more to an improvement in AUC than AN_{ch}. A potential explanation for this can be that, given the small number of churners (only 10.92% out of observed customer base), forcing walks to visit AN_{ch} results in less coverage of the network which impedes the learning process, as opposed to visiting AN_{rfm} nodes which allow for reaching nodes with different characteristics. To further test this hypothesis, we experimented with the more strict version of local meta path to completely forbid visiting AN_{rfm} nodes (hence leaving 0.01 probability for AN_{ch} but setting probability for AN_{rfm} to 0). This resulted in further deterioration of the AUC score to 0.6044 (and lift to 1.5742).

7 Conclusion

In this paper we make a synergy of existing advancements made in the direction of leveraging networked data for telco churn prediction and probabilistic meta paths-based representation learning. The contributions of this work are three-fold.

First, we raise a valid question of the nature of call graph extensions used for solving churn prediction in telco and shift the focus from a homogeneous to a heterogeneous perspective. Second, we devise probabilistic meta paths adapted for the data and problem at hand. Third, we perform experimental analysis and show the benefits of applying representation learning on graphs using probabilistic meta paths, both in terms of the improvement in the predictive results as well as in promising insights regarding the interplay of meta path type and predictive outcome.

A potential extension of this work can be the further refinement of node types in the obtained graphs that would consequently expand the space of potential meta paths. A more detailed study might be necessary to correlate different types of meta paths with social phenomena such as homophily and social influence as we see a lot of potential in this direction.

References

[Bengio *et al.*, 2003] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

[Bengio *et al.*, 2013] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, August 2013.

[Cao *et al.*, 2016] Shaosheng Cao, Wei Lu, and Qiongkai Xu. Deep neural networks for learning graph representations. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 1145–1152. AAAI Press, 2016.

[Castanedo *et al.*, 2014] F. Castanedo, G. Valverde, J. Zaratiegui, and A. Vazquez. Using deep learning to predict customer churn in a mobile telecommunication network. 2014.

[Chang *et al.*, 2015] Shiyu Chang, Wei Han, Jiliang Tang, Guo-Jun Qi, Charu C. Aggarwal, and Thomas S. Huang. Heterogeneous network embedding via deep architectures. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 119–128, New York, NY, USA, 2015. ACM.

[Chen and Sun, 2017] Ting Chen and Yizhou Sun. Task-guided and path-augmented heterogeneous network embedding for author identification. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, pages 295–304, New York, NY, USA, 2017. ACM.

[Chen *et al.*, 2012] Zhen-Yu Chen, Zhi-Ping Fan, and Minghe Sun. A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data. *European Journal of operational research*, 223(2):461–472, 2012.

[Cheng and Chen, 2009] Ching-Hsue Cheng and You-Shyang Chen. Classifying the segmentation of customer value via rfm model and rs theory. *Expert systems with applications*, 36(3):4176–4184, 2009.

[Dasgupta *et al.*, 2008] Koustuv Dasgupta, Rahul Singh, Balaji Viswanathan, Dipanjan Chakraborty, Sougata Mukherjea, Amit A Nanavati, and Anupam Joshi. Social ties and their relevance to churn in mobile telecom networks. In *Proceedings of the 11th conference on Extending database technology: Advances in database technology*, pages 668–677. ACM, 2008.

[Dong *et al.*, 2017] Yuxiao Dong, Nitesh V. Chawla, and Ananthram Swami. Metapath2vec: Scalable representation learning for heterogeneous networks. In *KDD*, 2017.

[Fu *et al.*, 2017] Tao-yang Fu, Wang-Chien Lee, and Zhen Lei. Hin2vec: Explore meta-paths in heterogeneous information networks for representation learning. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 1797–1806, New York, NY, USA, 2017. ACM.

[Grover and Leskovec, 2016] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 855–864. ACM, 2016.

[Huang *et al.*, 2015] Yiqing Huang, Fangzhou Zhu, Mingxuan Yuan, Ke Deng, Yanhua Li, Bing Ni, Wenyuan Dai, Qiang Yang, and Jia Zeng. Telco churn prediction with big data. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 607–618. ACM, 2015.

[Hughes, 1994] AM Hughes. *Strategic Database Marketing: The Master Plan for Starting and Managing a Profitable, Customer-Based Marketing program*. Probus Publishing Co., Chicago, IL., 1994.

[Kim *et al.*, 2014] Kyoungok Kim, Chi-Hyuk Jun, and Jaewook Lee. Improved churn prediction in telecommunication industry by analyzing a large network. *Expert Systems with Applications*, 41(15):6575–6584, 2014.

[McCarty and Hastak, 2007] John A McCarty and Manoj Hastak. Segmentation approaches in data-mining: A comparison of rfm, chaid, and logistic regression. *Journal of business research*, 60(6):656–662, 2007.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[Mitrović *et al.*, 2017] Sandra Mitrović, Gaurav Singh, Bart Baesens, Wilfried Lemahieu, and Jochen de Weerdt. Scalable rfm-enriched representation learning for churn prediction. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 79–88, Oct 2017.

[Modani *et al.*, 2013] Natwar Modani, Kuntal Dey, Ritesh Gupta, and Shantanu Godbole. Cdr analysis based telco churn prediction and customer behavior insights: A case study. In *International Conference on Web Information Systems Engineering*, pages 256–269. Springer, 2013.

[Perozzi *et al.*, 2014] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM, 2014.

[Sun and Han, 2013] Yizhou Sun and Jiawei Han. Mining heterogeneous information networks: A structural analysis approach. *SIGKDD Explor. Newsl.*, 14(2):20–28, April 2013.

[Sun *et al.*, 2011] Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S. Yu, and Tianyi Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. In *In VLDB 2011*, 2011.

[Tang *et al.*, 2015a] Jian Tang, Meng Qu, and Qiaozhu Mei. Pte: Predictive text embedding through large-scale heterogeneous text networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1165–1174, New York, NY, USA, 2015. ACM.

[Tang *et al.*, 2015b] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1067–1077. ACM, 2015.

[Umayaparvathi and Iyakutti, 2017] V. Umayaparvathi and K. Iyakutti. Automated feature selection and churn prediction using deep learning models. *International Research Journal of Engineering and Technology (IRJET)*, 4(3), 2017.

[Verbeke *et al.*, 2014] Wouter Verbeke, David Martens, and Bart Baesens. Social network analysis for customer churn prediction. *Applied Soft Computing*, 14:431–446, 2014.

[Wangperawong *et al.*, 2016] A. Wangperawong, C. Brun, O. Laudy, and R. Pavasuthipaisit. Churn analysis using deep convolutional neural networks and autoencoders. *CoRR*, abs/1604.05377, 2016.

[Zhang *et al.*, 2012] Xiaohang Zhang, Ji Zhu, Shuhua Xu, and Yan Wan. Predicting customer churn through interpersonal influence. *Knowledge-Based Systems*, 28:97–104, 2012.