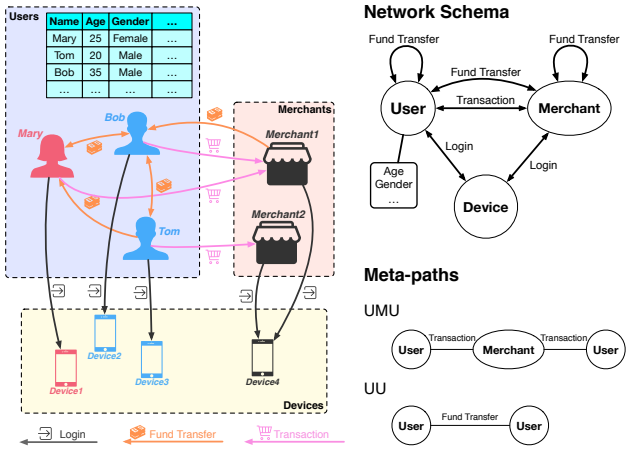
Cash-Out User Detection Based on Attributed Heterogeneous Information Network with a Hierarchical Attention Mechanism

Binbin Hu,[1] Zhiqiang Zhang,[2] Chuan Shi,[1] Jun Zhou,[2] Xiaolong Li,[2] Yuan Qi[2]
[1] Beijing University of Posts and Telecommunications
[2] AI Department, Ant Financial Services Group
{hubinbin, shichuan}@bupt.edu.cn, {lingyao.zzq, jun.zhoujun, xl.li, yuan.qi}@antfin.com

Abstract

As one of the major frauds in financial services, cash-out fraud is that users pursue cash gains with illegal or insincere means. Conventional solutions for the cash-out user detection are to perform subtle feature engineering for each user and then apply a classifier, such as GDBT and Neural Network. However, users in financial services have rich interaction relations, which are seldom fully exploited by conventional solutions. In this paper, with the real datasets in Ant Credit Pay of Ant Financial Services Group, we first study the cash-out user detection problem and propose a novel hierarchical attention mechanism based cash-out user detection model, called HACUD. Specifically, we model different types of objects and their rich attributes and interaction relations in the scenario of credit payment service with an Attributed Heterogeneous Information Network (AHIN). The HACUD model enhances feature representation of objects through meta-path based neighbors exploiting different aspects of structure information in AHIN. Furthermore, a hierarchical attention mechanism is elaborately designed to model user's preferences towards attributes and meta-paths. Experimental results on two real datasets show that the HACUD outperforms the state-of-the-art methods.



(a) Scenario of credit payment service (b) Network schema and meta-path examples

Figure 1: The AHIN of the scenario of credit payment service.

Introduction

Credit Payment Services, such as offline credit card services in commercial banks and online credit payments in internet financial institutions, are widely used in many aspects of daily life and bring convenience to both users and merchants. However, ever-increasing frauds have seriously influenced the security of credit payment services. *Cash-out* fraud is to pursue cash gains with illegal or insincere means, e.g., through buying pre-paid cards or other goods then reselling them. With the rapid development of e-commerce, it has become one of the major frauds on various kinds of credit payment services. Cash-out fraud behavior is illegal and may cause financial venture, since the probability of loan default is much higher for cash-out users in most cases. Therefore, *cash-out user detection* becomes one of the most important components of the fraud detection system in financial institutions.

The goal of cash-out user detection is to predict whether a user will do cash-out transactions or not in the future. Thus this problem can be formulated as a classification problem. Conventional solutions first perform subtle feature engineering for each user, and then a classifier, such as tree-based model or neural network, is trained based on these features. The key point of these methods is to extract statistical features of users from different aspects, such as user profile, credit history, transaction summarizing, and recent behaviors in other relative businesses. Conventional methods make prediction mainly based on the statistical features of a certain user, but seldom fully exploit the interaction relations between users, which may be beneficial to the cash-out user detection problem.

In fact, there are rich interaction relations in the scenario of credit payment service, which are really important to the cash-out user detection problem. Fig. 1a demonstrates a general scenario of credit payment service, where there are three types of objects: users, merchants, and devices (the way to access services, e.g., websites, desktops, mobile apps, wifi devices, etc.). Besides the attribute information, these objects also have rich interaction information *e.g.,* the fund transfer relation among users, the login relation between

users and devices, and the transaction relation between users and merchants. The cash-out users not only have abnormal features, but also behavior abnormally in interaction relations. For example, the cash-out users may simultaneously have many transaction and fund transfer interactions with particular merchants, which is hard to be exploited by traditional feature extraction.

In order to exploit the interaction relations and feature information, we propose to model the scenario of credit payment service with an ***A**ttributed **H**eterogeneous **I**nformation **N**etwork* (AHIN). The recently emerging ***H**eterogeneous **I**nformation **N**etwork* (HIN) (Shi et al. 2017), consisting of multiple types of nodes and links, has been proposed as a powerful information modeling method for characterizing data heterogeneity (Sun et al. 2011; Zhao et al. 2017). Furthermore, in order to incorporate the attribute information of objects, we extend traditional HIN to AHIN, where objects in HIN may contain attributes (or termed as features). Fig. 1b shows the network schema of the AHIN in the scenario of credit payment service, which clearly illustrates the objects and their interactions. Several efforts have been made for mining HIN and shown promising performance in various kinds of applications (Dong, Chawla, and Swami 2017; Sun and Han 2012; Shi et al. 2018). However, they are usually designed for specific task and only exploit structure information, so they cannot be directly applied for the AHIN and the cash-out user detection problem.

In this paper, we first study the cash-out detection problem under the AHIN framework, and propose a novel **H**ierarchical **A**ttention mechanism based **C**ash-out **U**ser **D**etection model, called HACUD. The basic idea of HACUD is to significantly enhance the feature representation of objects through fully exploiting interaction relations, i.e., with the help of meta-path based neighbors in AHIN. Inspired by (Kipf and Welling 2017; Zhang et al. 2018) and our observations on real data, we assume that the feature representation of objects, besides intrinsic features, are also constituted by the features of their neighbors. We propose the concept of meta-path based neighbors to exploit rich structure information in AHIN. That is, we can find neighbors of a node through the assigned meta-path (a relation sequence connecting two nodes). It has several advantages: (1) It can capture different aspects of structure information through different meta-paths (Han et al. 2018); (2) It greatly reduces the dimension of representation space, compared to traditional network representation learning methods; (3) It is potential to predict new-coming nodes dynamically. Furthermore, we assume that object attributes and meta-paths have different importances, and elaborately design a hierarchical attention mechanism to learn user preferences towards attributes and meta-paths. Specifically, the first layer of our attention mechanism models the user's attention in the feature space (*i.e.,* attributes), while the second layer captures the different contributions of different meta-paths for the prediction task. Finally, a cash-out probability is predicted based on aggregated feature representation with a multi-layer perceptron.

In summary, our work has the following contributions.

- We are the first to study the cash-out users detection problem, which is a very important and widely existing problem in financial fraud field.
- We propose to model the cash-out user detection problem as a classification problem in AHIN which is constituted by different types of objects and their rich interactions in the scenario of credit payment service.
- We propose a novel model HACUD to solve the problem, which employs meta-path based neighbors to fully exploit structure information and a hierarchical attention mechanism to automatically learn the importance of attributes and meta-paths.
- Extensive experiments on two real datasets illustrate the best performance of the proposed HACUD compared to the state of arts, as well as the benefits of hierarchical attention mechanism.

Preliminary

A HIN is a special kind of information network, which contains either multiple types of objects or multiple types of links (Sun and Han 2012). In order to integrate widely existing attribute information of objects, we further extend HIN to attributed heterogeneous information network (AHIN) as follows.

Definition 1 ***Attributed Heterogeneous Information Network** (AHIN). An AHIN is denoted as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathbf{X}\}$ consisting of an object set $\mathcal{V}$, a link set $\mathcal{E}$ and an attribute information matrix*[1] *$\mathbf{X} \in \mathbb{R}^{|\mathcal{V}| \times k}$. An AHIN is also associated with a node type mapping function $\phi : \mathcal{V} \rightarrow \mathcal{A}$ and a link type mapping function $\psi : \mathcal{E} \rightarrow \mathcal{R}$. $\mathcal{A}$ and $\mathcal{R}$ denote the sets of predefined object and link types, where $|\mathcal{A}| + |\mathcal{R}| > 2$.*

In AHINs, two objects can be connected via different semantic paths, which are called meta-paths.

Definition 2 ***Meta-path** (Sun et al. 2011). A meta-path ρ is defined as a path in the form of $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \cdots \xrightarrow{R_l} A_{l+1}$ (abbreviated as $A_1 A_2 \cdots A_{l+1}$), which describes a composite relation $R = R_1 \circ R_2 \circ \cdots \circ R_l$ between object A_1 and A_{l+1}, where $\circ$ denotes the composition operator on relations.*

Example 1 *As shown in Fig. 1a, we construct an AHIN to model the scenario of credit payment service in which cash-out fraud usually happens. It consists of multiple types of objects(*i.e., *User (U), Merchant (M), Device (D)) with rich attributes and relations (*i.e., *fund transfer relation between users and transaction relation between users and merchants). In the AHIN, two users can be connected via multiple meta-paths,* e.g., *"User-(fund transfer) -User" (UU) and "User-(transaction) -Merchant-(transaction)-User" (UMU). Different meta-path always convey different semantics. For example, the UU path connects users having fund transfer from one to another, while the UMU connects users having transactions with the same merchants.*

As a major technical approach, meta-path based data mining methods have been extensively studied in HINs (Shi et

[1] In our work, we discretize the original attributes to the same dimension.

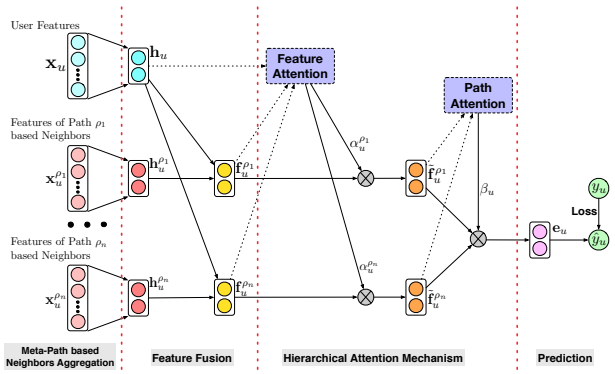
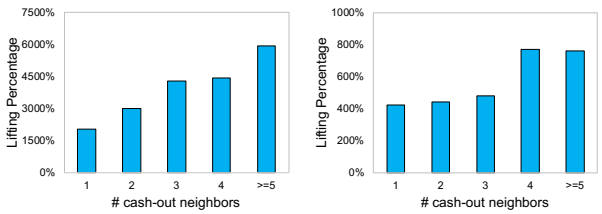


Figure 2: The architecture of the proposed model



(a) UMU (b) UU

Figure 3: The lifting percentages of cash-out rate in users with different amount of cash-out neighbors against users without any cash-out neighbor in two meta-paths.

al. 2017). Giving a meta-path ρ, there exists multiple specific neighbors *w.r.t.* each user. This neighbors set can reveal semantics and structure information of users in AHIN.

Definition 3 ***Meta-path based Neighbors***. *Giving a user u in an AHIN, the meta-path based neighbors is defined as the set of aggregate neighbors under the given meta-path for the user u in the AHIN.*

Example 2 *Take Fig. 1a as an example. Giving the meta-path UU, the neighbors of "Mary" are "Bob" and "Tom". Similarly, the neighbors of "Mary" based on meta-path UMU are "Tom". Obviously, meta-path based neighbors can exploit different aspects of structure information in AHIN.*

Here we model the scenario of credit payment service with an AHIN which can comprehensively integrate rich features and interaction information. Furthermore, we define the cash-out user detection problem under the AHIN framework as follows.

Definition 4 ***Cash-out User Detection Problem under AHIN***. *In the cash-out user detection problem, various kinds of objects and their interactions can be modeled as an AHIN $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathbf{X}\}$. In our setting, we focus on detecting cash-out users who are a subset of the node set, denoted as $\mathcal{U} \subset \mathcal{V}$. We assign a label $y_u \in \{0, 1\}$ on each user $u \in \mathcal{U}$ to indicate whether he/she is a cash-out user or not. Given the AHIN $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathbf{X}\}$ and the training set $\mathcal{D} = \{(u, y_u)\}$, the goal is to predict the cash-out probability p_{u^t} of user u^t in the test set.*

The Proposed Model

In this section, we firstly analyze the effect of meta-path based neighbors on the cash-out user detection on real data and then present the proposed ***H**ierarchical **A**ttention mechanism based **C**ash-out **U**ser **D**etection model*, called **HACUD** shortly. We show the overall architecture of the proposed model in Fig. 2. Firstly, we aggregate neighbors for each user based on different meta-paths to integrate multiple aspects of structure information in AHIN, and then we transform and fuse the original features for better representation learning. Considering that different features and meta-paths have different importances, we design a hierarchical attention mechanism to model user preferences towards features and meta-paths.

Observations in Real Data

Intuitively, the cash-out users tend to aggregate closely through different kinds of interactions. Taking the AHIN in Fig. 1a as an example, cash-out users tend to make more transactions with merchants which sell particular goods (e.g., pre-pard cards) or interact with more deceivers. In order to validate the aggregation of cash-out users with respect to different relations, we do experiments on the real dataset in Ant Credit Pay of Ant Financial Services Group (see Ten Days Dataset in Experiments).

We first collect the meta-path based neighbors of each user based on two meta-paths (UMU meaning users having transactions with the same merchants and UU meaning users having fund transfer from one to another). For each meta-path, we count the number of neighbors who are cash-out user (called cash-out neighbor), and divide all users into different groups with respect to the number of their cash-out neighbors. The cash-out rate (i.e., the proportion of cash-out users) is calculated in each group. The lifting percentages of cash-out rate in different user groups against users without any cash-out neighbor, with respect to two meta-paths, are presented in Fig. 3. We have the following observations.

(1) Users with higher cash-out rate tend to have more cash-out neighbors. This observation illustrates that meta-path based neighbors have consistent behaviors with the original user, which implies that the features of users can stem from that of their meta-path based neighbors.

(2) Different meta-path based neighbors have different impacts on users. In Fig. 3, two meta-paths yield different lifting percentages. It inspires us that different meta-paths have different importances on users, which can be captured by recent emerging attention mechanism.

Meta-path based Neighbors Aggregation

Inspired by recently emerging graph convolutional networks (Kipf and Welling 2017; Dai, Dai, and Song 2016) and the above observations on real data, we think that feature representations of objects, besides their intrinsic features, are also composed of the features of their neighbors.

Based on this idea, we aggregate meta-path based neighbors for each user. Specifically, similar to recent attributed network embedding (Liang et al. 2018; Zhang et al. 2018), we adopt to represent a node *w.r.t.* a certain meta-path via aggregating features of its neighbors rather than the one-hot representation of its neighbors. For each user u, we can obtain the *aggregated features based on meta-path* ρ as below:

$$\mathbf{x}_u^\rho = \sum_{j \in \mathcal{N}_u^\rho} w_{uj}^\rho * \mathbf{x}_j, \tag{1}$$

where $\mathcal{N}_u^\rho$ is the neighbors of node j based on meta-path ρ and $\mathbf{x}_j$ represents the attribute information vector associated with node j. The given link weight $w_{uj} > 0$ for weighted networks and $w_{uj} = 1$ for unweighted networks.

Feature Fusion

For each user u, we can obtain its own feature $\mathbf{x}_u$ as well as a set of its neighbor aggregation features based on multiple meta-paths $\{\mathbf{x}_u^\rho\}_{\rho \in \mathcal{P}}$ where $\mathcal{P}$ denotes the set of meta-paths. For better representation learning, we set up a *feature fusion* part to transform and fuse the original features.

Firstly, we project the original sparse features to the low-dimensional dense representations, and obtain the *latent representations* of user u and his/her neighbors based on different meta-paths (i.e., $\mathbf{h}_u$ and $\mathbf{h}_u^\rho$), respectively:

$$\mathbf{h}_u = \mathbf{W}\mathbf{x}_u + \mathbf{b}, \quad \mathbf{h}_u^\rho = \mathbf{W}^\rho \mathbf{x}_u^\rho + \mathbf{b}^\rho, \tag{2}$$

where $\mathbf{W}^* \in \mathbb{R}^{D \times d}$ and $\mathbf{b}^* \in \mathbb{R}^d$ are the weight matrix and bias vector, respectively. D is the dimension of original feature space[2] and d is the dimension of latent representations. Next, we fuse the latent representations of a user and his/her neighbors based on each meta-path and add a fully-connected layer for more complicated interaction. For a meta-path ρ, we formulate the above procedure and obtain the *fusional representation* $\mathbf{f}_u^\rho$ *w.r.t.* meta-path ρ as below,

$$\mathbf{f}_u^\rho = \text{ReLU}(\mathbf{W}_F^\rho g(\mathbf{h}_u, \mathbf{h}_u^\rho) + \mathbf{b}_F^\rho). \tag{3}$$

Here, $\mathbf{W}_F^\rho \in \mathbb{R}^{d \times 2d}$ and $\mathbf{b}_F^\rho \in \mathbb{R}^d$ represent the weight matrix and bias vector based on meta-path ρ, respectively. $g(\cdot, \cdot)$ is the fusion function, which can be concatenation, addition or element-wise product (In our implementation, $g(\cdot, \cdot)$ is concatenation).

Hierarchical Attention

Intuitively, different users are likely to have different preferences over the features based on different meta-paths as well as attribute information. Concretely, a user may place different importances to different-aspect features based on meta-paths. Moreover, features also have different importances for the prediction task. Due to the effectiveness of attention mechanism in various machine learning tasks (Hu et al. 2018; Cheng et al. 2018; You et al. 2016), we design a hierarchical attention mechanism to capture user preferences towards features and meta-paths.

[2]The original attributes are discretized to sparse D-dimensional feature as the model input

Feature Attention Since different features might not contribute to the prediction task equally, we learn the aspect-specific attention weights over features conditioned on the involved user based on each meta-path. Given the user latent representation $\mathbf{h}_u$ and latent representation of his/her neighbors $\mathbf{f}_u^\rho$ based on meta-path ρ , we adopt a two-layer neural network to implement the attention.

$$\boldsymbol{v}_u^\rho = \text{ReLU}(\mathbf{W}_f^1[\mathbf{h}_u; \mathbf{f}_u^\rho] + \mathbf{b}_f^1), \tag{4}$$

$$\boldsymbol{\alpha}_u^\rho = \text{ReLU}(\mathbf{W}_f^2 \boldsymbol{v}_u^\rho + \mathbf{b}_f^2), \tag{5}$$

where $\mathbf{W}_f^*$ and $\mathbf{b}_f^*$ denote the weight matrix and bias vector, respectively and $[\cdot;\cdot]$ represents the concatenation of two vectors. Following the standard setting of neural attention networks, we normalize the above attention scores with the softmax function to obtain the final attention weights.

$$\hat{\alpha}_{u,i}^\rho = \frac{\exp(\alpha_{u,i}^\rho)}{\sum_{j=1}^K \exp(\alpha_{u,j}^\rho)}. \tag{6}$$

Then, the final representation of user u *w.r.t.* a meta-path ρ can be computed as follows,

$$\widetilde{\mathbf{f}}_u^\rho = \hat{\boldsymbol{\alpha}}_u^\rho \bigodot \mathbf{f}_u^\rho, \tag{7}$$

where "$\bigodot$" denotes the element-wise product.

Path Attention Given a user, following the above steps, we could obtain multiple representations based on multiple meta-paths, which are expected to collaborate with each other for better prediction. Following (Qu et al. 2017), we learn the attention weights over different meta-paths for collaboration. Specifically, we define the attention weight of meta-path ρ for user u using a softmax unit as follows:

$$\beta_{u,\rho} = \frac{\exp(\mathbf{z}^{\rho\,\mathrm{T}} \cdot \widetilde{\mathbf{f}}_u^C)}{\sum_{\rho' \in \mathcal{P}} \exp(\mathbf{z}^{\rho'\,\mathrm{T}} \cdot \widetilde{\mathbf{f}}_u^C)}, \tag{8}$$

where $\mathbf{z}^\rho \in \mathbb{R}^{|\mathcal{P}| * d}$ is the attention vector for meta-path ρ and $\widetilde{\mathbf{f}}_u^C$ is the concatenation of user u's representations *w.r.t.* all meta-paths (*i.e.,* $\widetilde{\mathbf{f}}_u^\rho$). After obtaining the path attention scores $\beta_{u,\rho}$, the final representation aggregating all meta-paths is given as the following weighted sum form:

$$\mathbf{e}_u = \sum_{\rho \in \mathcal{P}} \beta_{u,\rho} * \widetilde{\mathbf{f}}_u^\rho, \tag{9}$$

where $\widetilde{\mathbf{f}}_u^\rho$ is the representation of neighbors for user u based on meta-path ρ in Eq. 7.

Model Learning

Since neural networks have shown strong ability in modeling the complex interactions (He et al. 2017), we feed the obtained final representation (*i.e.,* $\mathbf{e}_u$) into multiple fully connected neural networks as follows,

$$\mathbf{z}_u = \text{ReLU}(\mathbf{W}_L \cdots \text{ReLU}(\mathbf{W}_1 \mathbf{e}_u + \mathbf{b}_1) + \mathbf{b}_L), \tag{10}$$

where $\mathbf{W}_*$ and $\mathbf{b}_*$ respectively denote the weight matrix and the bias vector for each layer. The predicted cash-out probability is obtained via a regression layer with a sigmoid unit:

$$p_u = \text{sigmoid}(\mathbf{w}_p^T \mathbf{z}_u + b_p). \tag{11}$$

Here $\mathbf{w}_p$ and b_p are the weight vector and the bias, respectively. As our task is classification, we model the objective function with maximum likelihood estimation, which can be formulated as follows:

$$\mathcal{L}(\Theta) = \sum_{\langle u, y_u \rangle \in \mathcal{D}} (y_u \log(p_u) + (1 - y_u) \log(1 - p_u)) + \lambda ||\Theta||_2^2, \quad (12)$$

where y_u and p_u represent the ground truth and the predicted cash-out probability of user u, respectively. Θ is the parameter set of the proposed model and λ is the regularizer parameter. The stochastic gradient descent (SGD) or its variants are adopted for optimization.

Discussion

As mentioned above, the proposed model is a flexible framework to leverage structure and attribute information through capturing multiple aspects in AHIN. Based on different meta-paths, we can integrate various kinds of heterogeneous information to enhance the prediction performance. Moreover, we represent a user via the feature aggregation of his/her neighbors based on meta-paths, which is a natural way to combine network structure and attribute information. Compared to traditional network embedding methods (Wang, Cui, and Zhu 2016; Dai, Dai, and Song 2016), which represent nodes via their context (*e.g.,* adjacency matrix), our method is more suitable for large-scale networks. Specifically, the dimension of original input can be reduced from $\mathcal{O}(|\mathcal{V}|)$ to $\mathcal{O}(D)$, where $|\mathcal{V}|$ is the total number of nodes in network and D is the feature dimension after discretization for each user ($D \ll |\mathcal{V}|$). For a new user which never appears in training set, the proposed model can also learn the representation through his/her meta-path based neighbors in networks. Therefore, our model has the ability to give predictions dynamically to some extent.

Experiments

In this section, we construct the experimental evaluation and present the result analysis.

Evaluation Dataset

With the real-world datasets in Ant Credit Pay, an online credit payment service provided by Ant Financial Services Group, we extract two sub-datasets for the evaluation, namely Ten Days Dataset (contains 1.88 million users ranging from 2018/03/21 to 2018/03/31 for training) and One Month Dataset (contains 5.16 million users ranging from 2018/03/01 to 2018/03/31 for training). For both datasets, we predict the cash-out probability of users in 2018/05/01 (around 0.17 million users). In our datasets, we define the positive samples as users who have involved in suspected cash-out transactions within one month and the negative samples as users who have never involved in suspected cash-out transactions within one month. Since we utilize data in the next month when defining the label, the time interval between training and test set is one month. To be noted that since the cash-out fraud is very much in the minority among all transactions, the negative examples are sampled to keep the cash-out rate at around 2% in our datasets.

Table 1: Selected meta-paths and meta-path based neighbors statistics.

Meta-paths	#Neighbors (Min / Max / Avg.)
User-(transaction) -Merchant-(transaction)-User	1 / 16860 / 309
User-(fund transfer) -User	1 / 26235 / 150
User-(transaction) -Merchant	1 / 81 / 4

We construct an attributed heterogeneous information network based on the two datasets, consisting of 56.75 million users and 0.51 million merchants. In addition, the AHIN contains 77.40 million fund transfer relations between users and 20.64 million transaction relations between users and merchants. We extract 123 attributes for each user, including user profile, credit history, transaction summarizing, recent behaviors in other relative businesses and so on. Considering the scale of attribute value and the existence of missing value, we preproccess the two datasets with feature discretization.

Evaluation Metrics

We use the widely adopted metric to measure the performance of cash-out user detection , namely **AUC** (*i.e.,* Area Under the ROC Curve). The AUC metric is defined is:

$$AUC = \frac{\sum_{u \in \mathcal{U}^+} rank_u - \frac{|\mathcal{U}^+| \times (|\mathcal{U}^+|+1)}{2}}{|\mathcal{U}^+| \times |\mathcal{U}^-|}. \quad (13)$$

Here, $\mathcal{U}^+$ and $\mathcal{U}^-$ denotes the positive and negative set in the test set, respectively. And $rank_u$ indicates the rank of user u via the score of prediction.

Methods to Compare

We consider several representative methods for the cash-out user detection task, which can roughly be categorized into three type: (1) Attribute only or Structure only (GBDT, Node2vec, Metapath2vec). (2) Structure + Attribute (Node2vec + Feature, Metapath2vec + Feature). (3) Structure + Attribute + Label (Structure2vec, GBDT$_{Struct}$).

•**Node2vec** (Grover and Leskovec 2016) : It is a representation learning method on homogeneous network. We employ it to learn representations for nodes, ignoring node heterogeneity, and feed them into a classification model.

•**Metapath2vec** (Dong, Chawla, and Swami 2017) : It is a heterogeneous information network embedding method for learning node embedding with meta-path guided random walks. Similarly, we also feed the node representations into a classification model.

•**Node2vec + Feature** : We feed the features of node as well as the embeddings learned by Node2vec into a classification model.

•**Metapath2vec + Feature** : We feed the features of node as well as the embeddings learned by Metapath2vec into a classification model.

Table 2: Results of effectiveness experiments on two datasets *w.r.t.* the dimension of latent represantation d. A larger value indicates a better performance.

Algorithm	AUC							
	Ten Days Dataset				One Month Dataset			
	$d = 16$	$d = 32$	$d = 64$	$d = 128$	$d = 16$	$d = 32$	$d = 64$	$d = 128$
Node2vec	0.5893	0.5913	0.5926	0.5930	0.5980	0.6063	0.6009	0.6021
Metapath2vec	0.5914	0.5903	0.5917	0.5920	0.6005	0.5976	0.5995	0.5983
Node2vec + Feature	0.6455	0.6464	0.6510	0.6447	0.6541	0.6561	0.6607	0.6518
Metapath2vec + Feature	0.6456	0.6429	0.6469	0.6485	0.6550	0.6552	0.6523	0.6545
Structure2vec	0.6537	0.6556	0.6598	0.6545	0.6641	0.6632	0.6657	0.6678
GBDT	0.6389	0.6389	0.6389	0.6389	0.6467	0.6467	0.6467	0.6467
GBDT$_{Struct}$	0.6948	0.6948	0.6948	0.6948	0.6968	0.6968	0.6968	0.6968
HACUD	**0.7066**	**0.7115**	**0.7056**	**0.7049**	**0.7132**	**0.7160**	**0.7109**	**0.7154**

•**Structure2vec** (Dai, Dai, and Song 2016) : It is an effective approach for network embedding based on network structure and feature information with labeled data.

•**GBDT** (Friedman 2001) : It is a scalable tree-based model for feature learning and classification task. We feed node feature into GBDT.

•**GBDT**$_{Struct}$: Besides node feature, we also feed the aggregate features of meta-path based neighbors into GBDT.

Implementation Details

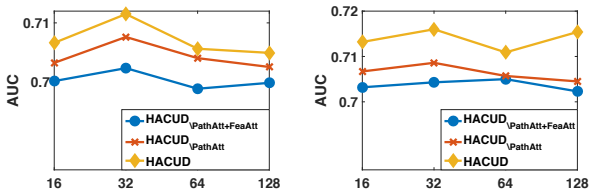
We implement the proposed model based on Tensorflow (Abadi et al. 2016). We utilize two hidden layers for prediction. We randomly initialize the model parameters with a xavier initializer (Glorot and Bengio 2010) and choose RMSProp (Tieleman and Hinton 2012) as the optimizer. Moreover, we set the batch size to 256, the learning rate to 0.002 and set the regularizer parameter $\lambda = 0.01$ to prevent overfitting. We report the selected meta-paths and meta-path based neighbors statistics information in Table 1. For the other comparison methods, we optimize their parameters according to literatures. Moreover, for all baselines, we implement them on parameter server based distributed learning systems(Zhou et al. 2017) for scaling up to large-scale datasets. And we select GBDT as the final classification model for the baselines.

Experimental Results

Performance Comparison. We report the comparison results of the proposed approach and baselines *w.r.t.* the dimension of latent representation d in Table 2. The major findings from the experimental results can be summarized as follows:

(1) Our model outperforms all the baselines, which indicates that our model adopts a more principled way to leverage interaction relations and attribute information for improving prediction performance. Our model achieves the best performance where the dimension of latent representation $d = 32$. And overall, the performance change trend is smooth, indicating that our model is not very sensitive to this parameter.

(2) Among these baselines, we can find that the overall performance order is as follows: (label + attribute + structure) based methods (*i.e.,* GBDT$_{Struct}$, Structure2vec) $>$ (attribute + structure) based methods (*i.e.,* Node2vec + Feature, Metapath2vec + Feature) $>$ structure or attribute only based method (*i.e.,* Node2vec, Metapath2vec, GBDT). It indicates that the better performances can be achieved through fusing more information. In addition, structure information (*i.e.,* interaction relations) is really helpful for performance improvement.

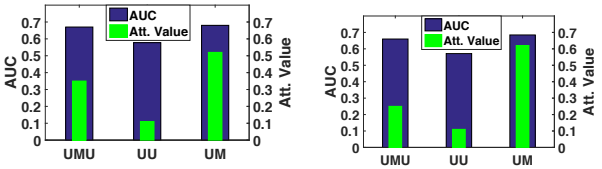


(a) Ten Days Dataset (b) One Month Dataset

Figure 4: Performance comparison of hierarchical attention *w.r.t.* the dimension of latent representation d.

(3) Compare the two variants of GBDT (*i.e.,* traditional GBDT and GBDT$_{Struct}$), we can find that GBDT$_{Struct}$ significantly outperforms traditional GBDT and other baselines, which further demonstrates the contribution of structural features provided by meta-path based neighbors in AHIN.

Effects of Hierarchical Attention. One of the major contributions of HACUD is hierarchical attention mechanism which learns the user preference towards features and meta-paths. In order to examine its effectiveness, we compare our model with its two variants, namely HACUD$_{\setminus PathAtt}$ (HACUD without path attention) and HACUD$_{\setminus PathAtt+FeaAtt}$ (HACUD without path and feature attention). For the performance comparison in Fig. 4, we can find that the overall performance order is as follows: HACUD $>$ HACUD$_{\setminus PathAtt}$ $>$ HACUD$_{\setminus PathAtt+FeaAtt}$. The results show that the hierarchical mechanism is able to better utilize the user feature and features generated by meta-paths in two aspects. First, different meta-paths have different contributions to cash-out user prediction, which cannot be treated equally (*i.e.,* HACUD$_{\setminus PathAtt}$). Second,



(a) Ten Days Dataset (b) One Month Dataset

Figure 5: Performances comparison on different meta-paths and corresponding attention values.

each user tends to place different importance to the various attributes for each meta-path. Ignoring such influence may not be able to achieve the promising performance for fully exploiting attribute and structure information (*i.e.,* HACUD$_{\backslash PathAtt+FeaAtt}$).

Furthermore, we report the performances based on single meta path and corresponding average attention value in Fig. 5. As we have observed, the performances of HACUD with different meta-paths and the corresponding attentions are positively correlated (*i.e.,* important meta-paths tend to attract more attentions). In other words, the proposed HACUD model is potential to let different users focus on the proper meta-paths.

Impact of Different Meta-paths. As mentioned above, our model utilizes a selected set of meta-paths. To further analyze the impact of different meta-paths, we gradually incorporate these meta-paths into our model and observe the performance change. In addition, we select GBDT as the baselines in this experiment. For convenience, we make the following denotation: (1) M_1 : user feature only; (2) M_2 : user feature + UMU; (3) M_3 : user feature + UMU + UU; (3) M_4 : user feature + UMU + UU + UM. As shown in the Fig. 6, we can observe that the performance would improve with the incorporation of more meta-paths, which demonstrates the effectiveness of structure information contained in different meta-paths. Specially, we can find that our model has a significant performance boost when adding the meta-paths UMU and UM. This finding is consistent with previous observation in Fig. 5, where these two meta-paths have better performances, accompanying with higher attention values.

Parameter Tuning Besides the dimension of latent representation d in Table 2, our model also involves another important tuning parameter λ in Eq. 12. We vary it in the set of $\{0.0, 0.0001, 0.001, 0.01, 0.1, 1.0\}$. As shown in Fig. 5, the optimal performance is obtained near $\lambda = 0.01$, indicating that λ cannot be set too small or too large to prevent overfitting and underfitting.

Related Work

As a newly emerging direction, heterogeneous information network (Shi et al. 2017) can model complex objects and their rich relations in real scenario. Due to the flexibility of HIN in modeling various kinds of heterogeneous data, many

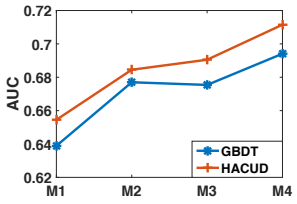


Figure 6: Impact of different meta-paths on Ten Days Dataset.

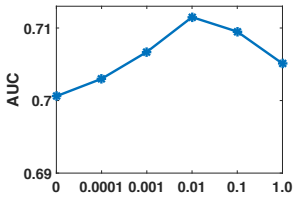


Figure 7: Impact of parameter λ on Ten Days Dataset.

meta-path based search and mining tasks have been explored in the past couple of years, including clustering (Sun et al. 2012), classification (Ji, Han, and Danilevsky 2011) and recommendation (Hu et al. 2018). Considering the plentiful attributes in the nodes, (Li et al. 2017) further proposes attributed heterogeneous information network to enrich objects' information content and study the problem of clustering objects in an AHIN. Traditional network mining methods do not pay much attention to node attribute information, which may play important roles in real applications. Therefore, we firstly propose to model the cash-out users detection problem as a classification problem in AHIN.

On the other hand, network embedding has shown its potential in structure feature extraction and has been successfully applied in many data mining tasks. Early network embedding methods focus on homogeneous network, which usually utilize network context information to represent nodes, *e.g.,* random walk based context (Perozzi, Al-Rfou, and Skiena 2014; Grover and Leskovec 2016), network neighborhood (Wang, Cui, and Zhu 2016; Tang et al. 2015), high order network proximity (Cao, Lu, and Xu 2015). Recently, attention is increasingly shifting towards heterogeneous network. (Dong, Chawla, and Swami 2017) obtain the context of nodes with meta-path based random walk and learn the HIN embedding through heterogeneous skip-gram model, while (Fu, Lee, and Lei 2017) capture rich relation semantics via neural network. Moreover, there are also several works attempting to fully analyze networks via embedding methods with features and labeled data, including GCN (Kipf and Welling 2017), Structure2vec (Dai, Dai, and Song 2016) and so on. Unfortunately, these methods are usually designed for specific task and only exploit partial information in networks, therefore they cannot be directly applied in the AHIN and the cash-out user detection problem for promising performance.

Conclusion

In this paper, we first study the cash-out user detection problem under the attributed heterogenous information network framework, constituted by objects and their relations in the scenario of credit payment service, and propose a novel HACUD model for the purpose. With the help of meta-path based neighbors, we aggregate features of objects from node attributes, as well as structure features generated by meta-paths. Furthermore, we design a hierarchical attention mechanism to model user preferences towards attributes and meta-paths. With the real datasets in Ant Credit Pay of Ant Financial Services Group, extensive experiments for the cash-user detection task demonstrate the effectiveness of our model. As future work, we will investigate to integrating more heterogeneous information (*e.g.,* interaction relations) and extending our model to semi-supervised scenario.

Acknowledgement

We would like to thank Feng Zhao, Yanming Fang and Quan Yu of Ant Financial Service Group. This work is supported in part by the National Natural Science Foundation of China (No. 61772082, 61702296, 61806020, 61375058), the National Key Research and Development Program of China (2017YFB0803304) and the Beijing Municipal Natural Science Foundation (4182043).

References

Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. 2016. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, 265–283.

Cao, S.; Lu, W.; and Xu, Q. 2015. Grarep: Learning graph representations with global structural information. In *CIKM*, 891–900.

Cheng, Z.; Ding, Y.; He, X.; Zhu, L.; Song, X.; and Kankanhalli, M. 2018. A3ncf: An adaptive aspect attention model for rating prediction. In *IJCAI*, 3748–3754.

Dai, H.; Dai, B.; and Song, L. 2016. Discriminative embeddings of latent variable models for structured data. In *ICML*, 2702–2711.

Dong, Y.; Chawla, N. V.; and Swami, A. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *SIGKDD*, 135–144.

Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* 1189–1232.

Fu, T. Y.; Lee, W. C.; and Lei, Z. 2017. Hin2vec: Explore meta-paths in heterogeneous information networks for representation learning. In *CIKM*, 1797–1806.

Glorot, X., and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 249–256.

Grover, A., and Leskovec, J. 2016. node2vec: Scalable feature learning for networks. In *SIGKDD*, 855–864.

Han, X.; Shi, C.; Wang, S.; Philip, S. Y.; and Song, L. 2018. Aspect-level deep collaborative filtering via heterogeneous information networks. In *IJCAI*, 3393–3399.

He, X.; Liao, L.; Zhang, H.; Nie, L.; Hu, X.; and Chua, T.-S. 2017. Neural collaborative filtering. In *WWW*, 173–182.

Hu, B.; Shi, C.; Zhao, W. X.; and Yu, P. S. 2018. Leveraging meta-path based context for top-n recommendation with a neural co-attention model. In *SIGKDD*, 1531–1540.

Ji, M.; Han, J.; and Danilevsky, M. 2011. Ranking-based classification of heterogeneous information networks. In *SIGKDD*, 1298–1306.

Kipf, T. N., and Welling, M. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR*.

Li, X.; Wu, Y.; Ester, M.; Kao, B.; Wang, X.; and Zheng, Y. 2017. Semi-supervised clustering in attributed heterogeneous information networks. In *WWW*, 1621–1629.

Liang, J.; Jacobs, P.; Sun, J.; and Parthasarathy, S. 2018. Semi-supervised embedding in attributed networks with outliers. In *SDM*, 153–161.

Perozzi, B.; Al-Rfou, R.; and Skiena, S. 2014. Deepwalk:online learning of social representations. In *SIGKDD*, 701–710.

Qu, M.; Tang, J.; Shang, J.; Ren, X.; Zhang, M.; and Han, J. 2017. An attention-based collaboration framework for multi-view network representation learning. In *CIKM*, 1767–1776.

Shi, C.; Li, Y.; Zhang, J.; Sun, Y.; and Philip, S. Y. 2017. A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering* 29(1):17–37.

Shi, C.; Hu, B.; Zhao, X.; and Yu, P. 2018. Heterogeneous information network embedding for recommendation. *IEEE Transactions on Knowledge and Data Engineering*.

Sun, Y., and Han, J. 2012. Mining heterogeneous information networks: principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery* 3(2):1–159.

Sun, Y.; Han, J.; Yan, X.; Yu, P. S.; and Wu, T. 2011. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment* 4(11):992–1003.

Sun, Y.; Norick, B.; Han, J.; Yan, X.; Yu, P. S.; and Yu, X. 2012. Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. In *SIGKDD*, 1348–1356.

Tang, J.; Qu, M.; Wang, M.; Zhang, M.; Yan, J.; and Mei, Q. 2015. Line:large-scale information network embedding. In *WWW*, 1067–1077.

Tieleman, T., and Hinton, G. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning* 4(2):26–31.

Wang, D.; Cui, P.; and Zhu, W. 2016. Structural deep network embedding. In *SIGKDD*, 1225–1234.

You, Q.; Jin, H.; Wang, Z.; Fang, C.; and Luo, J. 2016. Image captioning with semantic attention. In *CVPR*, 4651–4659.

Zhang, Z.; Yang, H.; Bu, J.; Zhou, S.; Yu, P.; Zhang, J.; Ester, M.; and Wang, C. 2018. Anrl: Attributed network representation learning via deep neural networks. In *IJCAI*, 3155–3161.

Zhao, H.; Yao, Q.; Li, J.; Song, Y.; and Lee, D. L. 2017. Meta-graph based recommendation fusion over heterogeneous information networks. In *SIGKDD*, 635–644. ACM.

Zhou, J.; Li, X.; Zhao, P.; Chen, C.; Li, L.; Yang, X.; Cui, Q.; Yu, J.; Chen, X.; Ding, Y.; et al. 2017. Kunpeng: Parameter server based distributed learning systems and its applications in alibaba and ant financial. In *SIGKDD*, 1693–1702.