Fraud Detection in Dynamic Interaction Network

Hao Lin, Guannan Liu, Junjie Wu, Yuan Zuo, Xin Wan, and Hong Li

Abstract—Fraud detection from massive user behaviors is often regarded as trying to find a needle in a haystack. In this paper, we suggest abnormal behavioral patterns can be better revealed if both sequential and interaction behaviors of users can be modeled simultaneously, which however has rarely been addressed in prior work. Along this line, we propose a COllective Sequence and INteraction (COSIN) model, in which the behavioral sequences and interactions between source and target users in a dynamic interaction network are modeled uniformly in a probabilistic graphical model. More specifically, the sequential schema is modeled with a hierarchical Hidden Markov Model, and meanwhile it is shifted to the interaction schema to generate the interaction counts through Poisson factorization. A hybrid Gibbs-Variational algorithm is then proposed for efficient parameter estimation of the COSIN model. We conduct extensive experiments on both synthetic and real-world telecom datasets in different scales, and the results show that the proposed model outperforms some competitive baseline methods and is scalable. A case is further presented to show the precious explainability of the model.

Index Terms—fraud detection, sequential schema, interaction schema, telecommunication network, probabilistic graphical model

1 Introduction

Fraudulent behaviors are always evolving with the goal for improper or illegal benefits, resulting in severe social issues due to the huge economic loss and resource waste. Fraudsters can hide easily for the anonymity of the Internet, and the situation becomes even worse in the big data era because fraudulent behaviors are usually blended in the large amount of normal behaviors—like a needle in a haystack. For example, telecommunication fraud is an emerging type of fraudulent behaviors (the fraudsters make a large amount of calls in order to induce the callees to transfer money to their accounts through a deceptive story) that are hidden inside the huge volumes of calling records. Similarly, in finance scenarios, money laundering can take place between bank accounts to transform illegal profits of crimes into legitimate assets, and is hard to reveal from the massive bank transferring records. In addition, fraudulent behaviors like shilling attacks can also arise from comments and likes in social networks, product reviews in e-commerce, etc., where we always encounter the extreme data imbalance problem.

In order to achieve abnormal profits, fraudsters have to behave differently from normal users in various perspectives. For instance, in telecom frauds, fraudsters may dial a large number of callees within a short time period and pretend to be an officer in the tax bureau accusing the callees of tax evasion, such that they can have higher possibility to obtain potential targets as victims. Each calling record may have no significant difference compared with a normal calling record, if checked individually; however, the consecutive dialing behaviors can be sharply different from normal calls in terms of some continuously calling strategy and/or some pre-selected called targets. Also, a transfer-out account may transfer a certain amount of money to a large number of different accounts consecutively for money laundering, which would also show distinctive behavioral patterns from the normal ones in terms of sequential transferring behaviors and the transferred target accounts.

Specifically, fraudulent behaviors are driven by some underlying mechanisms entailing two parties, i.e., the sources and the targets. These mechanisms can be revealed from the sequential actions taken by both sides, as well as the interactive structure formed during the interactions. Such structure is generally represented as a bipartite network, with the nodes representing the users and the edges characterizing the interactive levels. Then, the fraud detection problem can be reduced to detecting anomalies in a dynamic interaction network, where each individual has both independent and interactive behaviors driven by sequential and interaction schemas, respectively. Behaviors that deviate from the normal can be recognized as fraudulence.

Prior work usually tackles the fraud detection problem from the sequence or network schema separately. For example, [1] considers the problem of anomaly detection in heterogeneous, multivariate, variable-length time series datasets with a focus on the aviation safety domain, while some others use graph measures [2], [3] or latent factor models [4], [5] to uncover suspicious patterns from graph data. However, the interactive and sequential behaviors are

- *H. Lin, G. Liu, Y. Zuo, and H. Li are with the School of Economics and Management, Beihang University, Beijing 100191, China. E-mail: {linhao2014, liugn, zuoyuan, hong_lee}@buaa.edu.cn.*
- *J. Wu is with the School of Economics and Management, Beihang University, Beijing 100191, China, and also with the Beijing Advanced Innovation Center for Big Data and Brain Computing, and the Beijing Key Laboratory of Emergency Support Simulation Technologies for City Operations, Beihang University, Beijing 100191, China. E-mail: wujj@buaa.edu.cn.*
- *X. Wan is with the National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100191, China. E-mail: wanxin@cert.org.cn.*

Manuscript received 3 Apr. 2018; revised 19 Mar. 2019; accepted 9 Apr. 2019. Date of publication 23 Apr. 2019; date of current version 10 Sept. 2020. (Corresponding authors: Guannan Liu and Junjie Wu.) Recommended for acceptance by C. Ordonez. Digital Object Identifier no. 10.1109/TKDE.2019.2912817

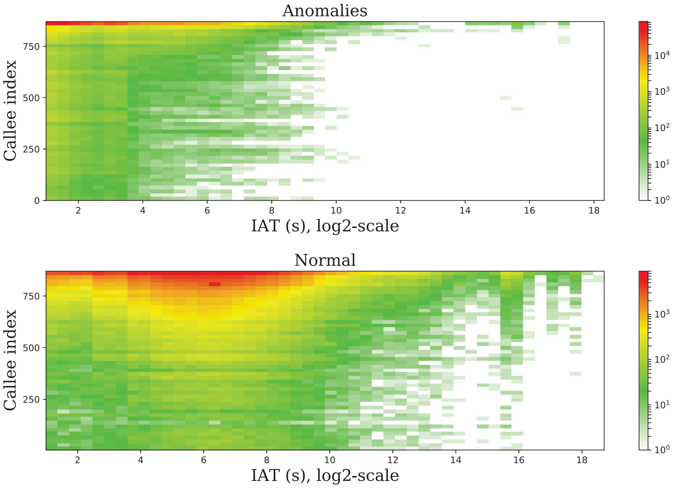


Fig. 1. Comparison of correlation matrix (sequential and interaction patterns) between fraudulent and normal callers. Vertical axis corresponds to the index of callees sorted by popularity, where a larger index indicates a more popular callee. Horizontal axis corresponds to the $\log_2$ value of the inter-arrival time (IAT) of sequential calls. Different colors represent different numbers of calls, the redder the larger.

essentially interweaved; that is, the connections between the two parties are established from the consecutive behaviors of both the sources and targets.

We take a small example to illustrate the synergistic effect. We observe the correlation between the sequences and the interaction structures from a telecom calling records dataset obtained from a telecom operator of China. Specifically, we collect the calling sequences of each caller and compute the time intervals between consecutive calls. Meanwhile, we regard the number of calls to a callee as her popularity and sort them accordingly. We then obtain the correlation matrices between the calling interval and the calling target for the normal and fraudulent source users, respectively. As shown in Fig. 1, the calling behavioral patterns of the normal and fraudulent sources are dramatically different in the two matrices. The fraudulent users seem concentrated on a small group of popular callees with a rather short interval, while the calling records of normal callers distribute much more dispersive, with quite a few having longer calls to less popular callees. These different behavioral patterns, however, cannot be obviously manifested if viewed only from single perspective. For example, considering only the calling intervals, both normal and fraudulent callers have considerable amount of calling behaviors within short intervals such that the frauds cannot be easily detected. If we collectively take the calling targets into account, the calling patterns when the intervals are small can be notably distinguished because the fraudulent users just call the most vulnerable targets, while in contrast, normal users would interact with more diversified target users with varied popularities.

Therefore, in order to detect fraudulent behaviors in a dynamic interaction network more accurately, we need to model the sequential behaviors and the corresponding interaction structures embedded in the behaviors collectively. To that end, in this paper, we propose a COllective Sequence and INteraction (COSIN) model to capture the synergistic effect of user behaviors, which is among the earliest to simultaneously model different types of behaviors for fraud detection. Considering the time-variant and individual-oriented nature of sequential behaviors, we introduce a hierarchical Hidden Markov Model (HMM) to capture the sequential schema-dependent transitional patterns. Specifically, a latent sequential schema is chosen for each time slot, which further generates the latent mode and the observed sequential behaviors consecutively in a probabilistic model. Moreover, the sequential schema of each individual can be shifted to the interaction schema with a shift distribution, and the interaction schema of both parities together generates the interaction count in order to reveal the interactive patterns through a latent factor model. We conduct extensive experiments on both synthetic and real-world telecommunication datasets, and the results validate the effectiveness of our model in different fraudulent scenarios and with abundant competing methods. A case study is also presented to illustrate the explanability of the model, with the transitional modes indicating meaningful sequential patterns.

The remainder of this paper is organized as follows. Section 2 introduces related work. Section 3 details the model and the inference method. Experimental results are given in Section 4, and we finally draw conclusions in Section 5.

2 Related Work

The basic principle in fraud detection is to detect abnormal behaviors that deviate from normal behaviors. Considering the above-mentioned formation of interaction network, frauds can be detected from multiple perspectives including sequences, interactions, as well as underlying grouped behaviors. Prior work usually tackles fraud detection problems from one single perspective, which is summarized into three main streams as follows.

2.1 Sequence-Based Anomaly Detection

Sequence-based anomaly detection generally identifies an entire sequence to be anomalous if it deviates significantly from normal sequences [6]. The most intuitive methods are similarity-based techniques, which compute pairwise similarity between sequences using a specific similarity measure [7], [8]. Another widely used family of approaches are Markov models-based techniques, which model the generative process of sequence data from probabilistic perspective. [9] proposes to mutually enhance the two intertwined tasks (i.e., user grouping and mobility modeling) based on an ensemble of Hidden Markov Models [10], and can be further applied to anomalous movement detection. Melnyk et al. [1] jointly model the heterogeneous, multivariate continuous system dynamics including the sensor measurements of flights and duration of flight mode by combining vector autoregressive model (VAR) and semi-Markov model (SMM) [11], and anomalous flight segments can then be detected by measuring dissimilarities between the prediction and observation. vanilla HMM models the transitions between various hidden states through a unified transition matrix, which however might encounter problem in handling varied transitional patterns across different user groups and time.

2.2 Interaction-Based Anomaly Detection

Tremendous efforts have also been made to solve anomaly detection from the perspective of interactions between entities. One category employs graph-based techniques for spotting outliers [2], [3], [12], [13] since graphs can effectively capture complex correlations among inter-dependent data points and abnormal behaviors can manifest. Bipartite network is a particular type of graphs that models the interaction behaviors between two parties directly, where several studies have proposed to detect anomaly by employing the characteristics of bipartite network. For instance, [14] proposes to rank vertices of a bipartite graph based on the link structure and prior information about vertices, which can be further used for outlier detection. [15] constructs a bipartite graph for auction users and transactions, and employs a Markov random field model to detect suspicious patterns. Particularly, several prior studies on detecting telecommunication frauds also employ the interaction network structure to enhance the detection performance [16], [17], [18].

The second category applies latent factor models for exploiting latent suspicious structures. [5] assumes that each node has time-evolving roles, where the roles are low-rank latent representations extracted by using Non-negative Matrix Factorization (NMF) on node-feature matrices. In [4], Singular Value Decomposition (SVD) is used to decompose the correlation matrix constructed from pairwise nodes features at a particular time window, then anomalous time window can be recognized if behavioral patterns deviate significantly from the recent past behaviors. Besides, tensor analysis can be applied for spotting high-order anomalous patterns [19], [20], [21], [22].

There also exist a stream of studies [23] that focus on detecting anomalous structure in dynamic networks. In these studies, they generally define a scoring function for the normality of nodes, edges, or subgraphs [4], [24], [25], and then detect the abrupt changes of the scoring values with respect to the whole sequences. These methods generally require longer sequences, in which the behaviors remain relatively stable at most time and show abnormality in a particular time window. This burst detection scheme, however, is quite different from our detection philosophy.

2.3 Probabilistic-Based Method for Anomaly Detection

Probabilistic-based methods have also been extensively studied for the purpose of anomaly detection in recent years. Latent Dirichlet Allocation (LDA) is employed for user profiling in telecommunications, while a significant deviation from the normal activity of an individual user is assumed to be fraudulence [26]. [27] also resorts to LDA for user profiling, followed by threshold-type classification using the KL-divergence for anomaly detection. Xiong et al. [28], [29] propose a family of hierarchical probabilistic models based on Latent Dirichlet Allocation [30] for group anomaly detection. In particular, they propose Mixture of Gaussian Mixture Models (MGMM) [28] by assuming that each data point belongs to one group and a group of data points can be represented by a hierarchical mixture model. Also, Flexible Genre Model (FGM) in [29] further extends MGMM by using a flexible structure based on “genres” to model more complex normal behaviors. Moreover, Soleiman et al. [31] propose an anomalous topic model for detecting anomalous clusters in high-dimensional discrete data. Generally speaking, these probabilistic-based approaches are closely related to our work because they share the common purpose in modeling complex normal patterns to reveal anomalous patterns underlying the observed behaviors. But none of them has taken both sequential and interaction perspectives into account collectively.

More recently, Yu et al. [32] address that the group membership of each data point was not known in prior in real-world applications and propose a hierarchical Bayesian model GLAD to jointly learn the group membership and score the anomalous groups, in which they take both behavioral sequences and interactions into account. However, GLAD uses bag-of-words features for modeling point-wise behavioral sequences and thus ignores the consecutive dependencies of sequential data, making it incapable of capturing complex sequential transition patterns.

3 COLLECTIVE SEQUENCE AND INTERACTION

Many real-world human interactive activities involve two parties, i.e., source users that provoke interactive behaviors and target users that passively receive the behaviors, which forms an interaction network. For example, in telecommunication network, each calling record is initiated by a caller and received by a callee, which can be denoted as source and target user, respectively. Fraudulent behaviors can be blended and hidden in such massive interactive behaviors, in which source users that launch the potential fraudulence can be identified as fraudulent users while the target are recognized as victims. Similarly, in finance scenario, bank accounts can form an interaction network driven by money transfer events, with the sources denoted by the transfer-out accounts, and the targets are the transfer-in accounts.

It is worth noting that each individual may play both roles simultaneously as a source or a target user, and we can inspect the behaviors respectively by assuming the two types of behaviors to be independent of each other. Taking the telecommunication network as an example, though callers and callees are tightly coupled in calling records, their sequential behaviors can still be treated independently when we need to represent the sequential schema respectively. In essential, the proposed COSIN does not directly cope with each calling record between caller and callee, but handles the interaction count between the two parties, which can be viewed as a synergistic effect of the sequential schema of the two parties and the assumption of independence is thus reasonable.

More formally, we can define a bipartite network that consists of both sources and targets, with the edges between the parties indicating the interactive behaviors. The edges in the bipartite network are indeed formed at different time periods dynamically, driven by interactive behaviors between source and target users. Thus, each individual in the network can entail a behavioral sequence by ordering their interactive events with others, which is assumed to be independent of others.

As discussed previously, both the interactions and sequential behaviors can be collectively perceived to identify frauds, which can be revealed from a dynamic interaction network. Specifically, assume that we have a dynamic interaction network $\mathcal{G} = < S, T; \mathbf{Y}; \mathbf{D}, \mathbf{D}' >$, where S and T denote the set of source and target users respectively, $\mathbf{Y}$ represents the interaction matrix, and $\mathbf{D}, \mathbf{D}'$ are the sets of behavioral sequences of each node in the network. Then, given a source node $n \in S$ and target node $m \in T$, the interaction event $y_{n,m} \in \mathbf{Y}$ can be a real value in proportion to the number of interactive events between them. The behavioral sequences for the source node n and target node m are denoted as $\boldsymbol{\delta}_n \in \mathbf{D}$ and $\boldsymbol{\delta'}_m \in \mathbf{D}'$, respectively. Therefore, our goal is to collectively model the interactions and sequential behaviors of both the sources and targets from $\mathcal{G}$, and then frauds can be identified as those that deviate from normal behaviors.

3.1 General Ideas

We propose to collectively model the sequential behaviors of both sources and targets, as well as their interaction structure in a unified probabilistic framework, which we name as the COllective Sequence and INteraction model. To be specific, we model the behaviors in dynamic interaction network under two different schemas, i.e., *sequential schema* and *interaction schema*. In what follows, we first define the two types of behavioral schemas in terms of the behavioral sequences and the interaction network, respectively.

Definition 1 (Sequential schema). *The patterns that govern how users behave sequentially in the interaction network, or how the interactive behaviors between sources and targets are generated as a sequence.*

Definition 2 (Interaction schema). *The patterns that govern how sources establish connections with targets to form a bipartite interaction network.*

With respect to the two types of behavioral schemas, we propose to model the sequences and interactions of each individual, by introducing distributions over latent variables as well as lower-rank latent factors, respectively. Along this line, both the observed behaviors in the sequences and the interactions can be generated in a unified probabilistic framework. With regards to the sequential schema, we introduce a discrete distribution $\boldsymbol{\pi}_n$ over different latent schemas g for each source user $n \in S$, and meanwhile each user is also endowed with a latent factor $\boldsymbol{\theta}_n$ to represent the patterns that how the user interacts with the target users. Moreover, in order to bridge the gap between the sequential schema and the interaction schema, we propose a latent offset variable ϵ_n to align the two schemas $\boldsymbol{\pi}_n$ and θ_n.

3.2 Sequential Schema

The behavior of an individual at each time slot in the sequence can be represented as multi-dimensional attributes. For example, in the telecom scenarios, the calling time, duration time, the calling regions, *etc.*, are the available characteristics to describe behaviors. Among various behavioral characteristics, the time intervals between two consecutive behaviors are one of the strongest indicators for potential fraudulent behaviors, because they can directly disclose distinctive behavioral strategies for cheating purposes. Therefore in this paper, we focus on the time intervals between consecutive behaviors, i.e., for each user n, the observed interval sequence is $\boldsymbol{\delta}_n = (\delta_{1,n}, \delta_{2,n}, \ldots, \delta_{I_n,n})^\top$, $\forall\ \delta_{i,n} \in \mathbb{R}^+$. As a matter of fact, we can also take any type of behavioral characteristics other than the time intervals to organize the attributed behavioral sequences.

Hidden Markov Model is extensively studied for modeling sequences such as speech, trajectories, *etc.*, in which a latent state is assumed to generate the observed values at each time slot. We define the latent state as *mode*, and the observed time intervals between the consecutive behaviors can be viewed as a distribution over these discrete modes. Thus, a latent mode $z_{i,n} \in \{1, 2, \cdots, K\}$ is generated for each user n's behavior at time slot i, and then the time interval $\delta_{i,n}$ is generated from the mode specific distribution $\delta_{i,n} \sim p(\delta_{i,n}|z_{i,n})$, which is called *emission probability distribution*. In vanilla HMM [33], all the sequences share a uniform transition matrix $\mathbf{A} = \{a_{k',k}\}$, where $a_{k',k}$ denotes the probability of mode k' transiting to mode k. Thus the behavioral mode $z_{i,n}$ at time i is generated from $z_{i,n} \sim p(z_{i,n}|z_{i-1,n}; \mathbf{A})$.

However, transitional patterns may differ across different groups of users, which cannot be well captured by a uniform transition matrix with vanilla HMM. For example, business users may follow more stable transitional patterns of time intervals, while students may have more irregular transitional calling intervals such as "long → short", or "short → long". Also, the transitional patterns of fraudulent users may change more dramatically than normal users for the cheating purpose. In addition, the schemas governing the mode transition patterns may also be time-variate, such that each individual can switch their sequential schemas at each time.

Therefore, at each time slot i, besides the latent mode, we further introduce a latent variable $g_{i,n}$ to represent the particular sequential schema for user n. Given the schema g, we also introduce schema-dependent transitional matrix $\mathbf{A}_g$ to govern the mode transitions. Here, $\mathbf{A}$ is indeed a three-dimensional $G \times K \times K$ matrix. With regards to the specific sequential schema, the mode is generated from the probability distribution that depends on both schemas and the mode of the previous time slot, i.e., $z_{i,n} \sim p(z_{i,n}|z_{i-1,n}, g_{i,n}; \mathbf{A}_{g_{i,n}})$.

As shown in Fig. 2, the observed sequences are generated in a three-layered graphical model in spirit of HMM. Each user is governed by a discrete distribution π to choose the sequential schema at each time slot, which is generated from a Dirichlet prior $\boldsymbol{\alpha}$. For simplicity, we assume symmetric Dirichlet distribution, where all of the elements making up the Dirichlet prior, i.e., each entry of the concentration hyperparameter vector $\boldsymbol{\alpha}$, has the same value. We also assume that the transitional matrix $\mathbf{A}_g$ is generated from a symmetric Dirichlet distribution, where $\boldsymbol{d}$ is the concentration hyperparameter vector. The interval time $\delta_{i,n}$ for user n at time slot t is generated from an exponential distribution, in which the rate parameter $\mathbf{B}_k$ depends on the calling mode $z_{i,n} = k$. Then, the intrinsic generative process follows a path from the latent sequential schema to the latent mode, and finally to the observed values. More formally, the generative process of the observed sequences of each individual can be depicted as follows.

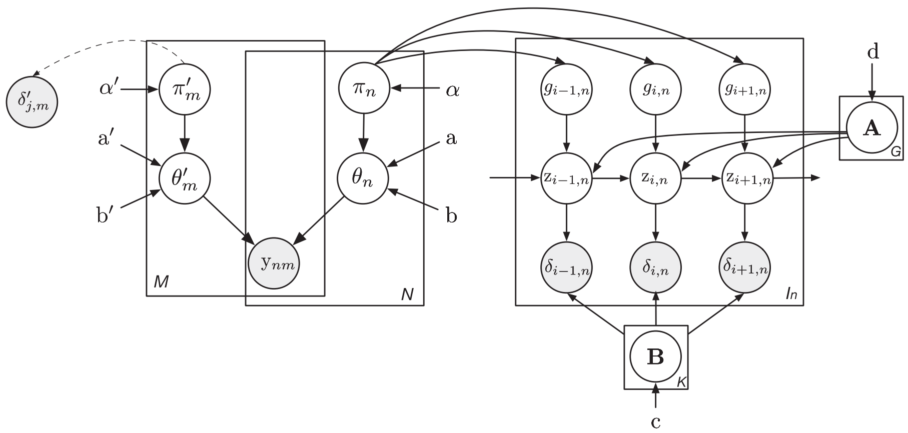


Fig. 2. The graphical representation of COSIN. Since the observed sequential intervals $\boldsymbol{\delta}'_{j,m}$ of the target users m are generated in the same process as the source user n, we replace their graphical representations through a dashed arrow for brevity.

1) For each source/target user n, draw the latent sequential schema distribution $\boldsymbol{\pi}_n \sim \mathrm{Dir}(\boldsymbol{\alpha})$.
2) For each pair of sequential schema and mode (g, k),
   a) Draw a K-dimensional transitional vector $\mathbf{A}_{g,k} \sim \mathrm{Dir}(\boldsymbol{d})$.
3) For each source user n,
   a) Draw a random sequence length variable I_n from a Poisson distribution.
   b) For each time slot $i \in 1, 2, \cdots, I_n$,
      i) Draw a discrete sequential schema indicator $g_{i,n} \sim \mathrm{Multinomial}(\boldsymbol{\pi}_n)$.
      ii) Draw a discrete mode indicator $z_{i,n} \sim \mathrm{Multinomial}(\mathbf{A}_{g_{i,n}, z_{i-1,n}})$.
      iii) Draw the interval time $\delta_{i,n} \sim \mathrm{Exp}(\mathbf{B}_{z_{i,n}})$.

3.3 Interaction Schema

The interaction patterns can indeed be revealed from the bipartite network formed between source and target users. Indeed, not only the binary interaction structure, but also the interaction intensity between the two parties, i.e., the number of interactions of a node pair, can further help disclose the underlying interaction schemas. For instance, fraudulent users may provoke large sum of calls toward a wide range of targets with low intensity; while normal users only make calls toward a small group of familiar contacts with higher interaction frequency. Particularly, we assume there are N sources and M targets in the interaction network, and let $\mathbf{Y} = \{y_{n,m}\}$, where $y_{n,m} \in \mathbb{N}$ indicates the number of interactions between source n and target m.

Apparently, latent factor model can be introduced to derive the latent representations of the interaction schema of both sources and targets by approximating the interaction count matrix. Particularly, the source n and the target m are represented by non-negative G-dimensional latent factors, i.e., $\boldsymbol{\theta}_n = (\theta_{1,n}, \theta_{2,n}, \ldots, \theta_{G,n})^\top$ and $\boldsymbol{\theta'}_m = (\theta'_{1,m}, \theta'_{2,m}, \ldots, \theta'_{G,m})^\top$, respectively, then the interaction count can be approximated by $y_{n,m} \approx \boldsymbol{\theta}_n^\top \cdot \boldsymbol{\theta'}_m$.

However, in reality, the count matrix can be rather sparse with only a small proportion of entries having positive values. Thus, we adopt the Poisson factorization [34] to resolve the sparsity rooted in the interaction matrix. Then, the observed interaction count $y_{n,m}$ is assumed to be generated by a Poisson distribution with its expectation given by the inner product of the latent variables:

$$y_{n,m} \sim \mathrm{Poisson}(\boldsymbol{\theta}_n^\top \cdot \boldsymbol{\theta'}_m). \tag{3}$$

As suggested by [34], an exponentially shaped prior Gamma distribution is capable of modeling the skewness in user activity and item popularity, which enforces sparsity in recommendation tasks. Similarly, to tackle the sparsity problem, we place Gamma priors on the latent factors as follows:

$$\begin{aligned} \boldsymbol{\theta}_n &\sim \mathrm{Gamma}(\theta^{(s)}, \theta^{(r)}), \\ \boldsymbol{\theta'}_m &\sim \mathrm{Gamma}(\theta'^{(s)}, \theta'^{(r)}). \end{aligned} \tag{4}$$

Let $b = \theta^{(r)}$, $b' = \theta'^{(r)}$ denote the rate priors, while $\theta^{(s)}$, $\theta'^{(s)}$ denote the shape priors.

3.4 Collective Model

The two different types of schemas, i.e., the sequential schema and the interaction schema, can be modeled separately for the corresponding behaviors in the dynamic interaction network. However, some fraudulent behaviors cannot be disclosed in one singe schema because different types may be correlated and can only be discovered when considering both schemas as a whole. For example, in order to discover anomalous behaviors of calling a large number of target users within a short time interval, we need to take a combined view to model the sequential and interaction schemas collectively. Therefore, in this paper, we integrate the two types of schemas in a unified model in order to uncover frauds from the collective perspectives.

As introduced previously, each user n can be represented by a latent discrete distribution $\boldsymbol{\pi}_n$ over sequential schemas, and can also be recognized by a latent factor $\boldsymbol{\theta}_n$ to reveal the interaction patterns. However, these two behavioral schemas are learned in distinctive latent space, which should be aligned by enforcing them to a common latent space. Specifically, a latent offset vector ϵ_n is introduced to align the sequential schema $\boldsymbol{\pi}_n$ and interaction schema $\boldsymbol{\theta}_n$, in which $\boldsymbol{\pi}_n$ can be viewed as prior information to be fed into the

TABLE 1
Gibbs Sampling Equations for Latent Schema and Mode

$$p(g_{i,n} = g|Rest) \propto \frac{\sum_{t=1,t\neq i}^{I_n} \Delta(g_{t,n} = g) + \alpha}{I_n - 1 + G\alpha} \cdot \frac{C_{g_{i,n},z_{i-1,n},k}^{-(i,n)} + d}{C_{g_{i,n},z_{i-1,n}}^{-(i,n)} + Kd} \tag{1}$$

$$\begin{aligned} p(z_{i,n} = k|Rest) \propto & \frac{c_1 + \sum_{n=1}^{N} \sum_{t=1,t\neq i}^{I_n} \Delta(z_{t,n} = k)}{c_2 + \sum_{n=1}^{N} \sum_{t=1,t\neq i}^{I_n} \delta_{t,n}\Delta(z_{t,n} = k)} \cdot \frac{C_{g_{i,n},z_{i-1,n},k}^{-(i,n)} + d}{C_{g_{i,n},z_{i-1,n}}^{-(i,n)} + Kd} \\ & \cdot \frac{C_{g_{i+1,n},k,z_{i+1,n}}^{-(i,n)} + d + \Delta(z_{i-1,n} = z_{i+1,n} = k)\Delta(g_{i+1,n} = g_{i,n})}{C_{g_{i+1,n},k}^{-(i,n)} + Kd + \Delta(z_{i-1,n} = k)\Delta(g_{i+1,n} = g_{i,n})}. \end{aligned} \tag{2}$$

Note: The conditional distribution for calling mode $z_{i,n}$ and schema indicator $g_{i,n}$ used in pointwise collapsed Gibbs sampler. Here, K is the number of latent modes, G is the number of sequential schemas, c_1 and c_2 are the shape parameter and rate parameter of the prior distribution of B_k respectively, $\Delta(\cdot)$ is the standard indicator function (i.e., equal to one if its argument is true and zero otherwise), $C_{g_{\cdot,n},z_{\cdot,n},k}^{-(i,n)}$ denotes the count of triplet $(g_{\cdot,n}, z_{\cdot,n}, k)$ appearing in source n's sequence but does not include the time step i, and $C_{g_{\cdot,n},z_{\cdot,n}}^{-(i,n)}$ denotes the count of the tuple $(g_{\cdot,n}, z_{\cdot,n})$ appearing in source n's sequence but does not include i.

latent factor $\boldsymbol{\theta}_n$. Thus, for each concrete sequential schema g, the latent offset latent factor $\epsilon_{n,g}$ can be generated from a Gamma distribution as follows:

$$\epsilon_{g,n} \sim \text{Gamma}(a, b), \tag{5}$$

where a and b serve as the two hyper-parameters for the Gamma distribution.

Then, the sequential schema $\boldsymbol{\pi}_{g,n}$ can be fed into the interaction schema $\boldsymbol{\theta}_{g,n}$ by a Gamma distribution $\text{Gamma}(\pi_{g,n}, b)$:

$$\begin{aligned} \boldsymbol{\theta}_{g,n} &\sim \text{Gamma}(\pi_{g,n}, b) + \epsilon_{g,n} \\ &\sim \text{Gamma}(\boldsymbol{\pi}_{g,n} + a, b). \end{aligned} \tag{6}$$

We can see from Eq. (4) that the shape prior for θ_n is $\theta^{\text{S}}_{g,n} = \pi_{g,n} + a$, where the hyper-parameter a can balance the effect that π_n placed over θ_n.

3.5 Parameter Learning

Given multiple users' calling sequences $\{\boldsymbol{\delta}_n\}_{n=1}^{N}$, $\{\boldsymbol{\delta'}_m\}_{m=1}^{M}$, and interaction count matrix $\mathbf{Y}^{N\times M}$, our goal is to infer emission distribution parameters $\{\mathbf{B}_k\}_{k=1}^{K}$, $\{\mathbf{B'}_k\}_{k=1}^{K}$, transition matrices $\{\mathbf{A}_g\}_{g=1}^{G}$, $\{\mathbf{A'}_g\}_{g=1}^{G}$, user latent representation vectors $\{\boldsymbol{\pi}_n\}_{n=1}^{N}$, $\{\boldsymbol{\pi'}_m\}_{m=1}^{M}$, and latent factors $\{\boldsymbol{\theta}_n\}_{n=1}^{N}$, $\{\boldsymbol{\theta'}_m\}_{m=1}^{M}$.

Due to the alignment for the two schemas, the exact posterior distribution $p(\Pi, \Pi', \mathbf{A}, \mathbf{A}', \mathbf{B}, \mathbf{B}', \Theta, \Theta'|Y, \Delta)$ is indeed intractable. Thus, we propose a hybrid Gibbs-Variational (HGV) algorithm for parameter estimation of the COSIN model. Specifically, we first develop a collapsed Gibbs sampling algorithm to infer the parameters for the hierarchical HMM component, then use variational methods to estimate the parameters related with Poisson factorization. The idea behind HGV is quite straightforward: 1) Collapsed Gibbs sampling does not assume any independence between parameters and hidden variables, and thus is expected to draw samples from the true posterior [35]. This makes our model more accurate to capture the complex dependencies between the hierachical consecutive hidden variables compared to other approximate inference methods; 2) Mean-field variational inference is practical and scalable to handle large scale interaction networks, whose effectiveness and efficiency have already been empirically demonstrated in [34].

3.5.1 Gibbs Sampler

As a typical kind of Markov Chain Monte Carlo (MCMC) algorithm, Gibbs sampling is adopted to produce a stream of samples from the posterior distributions of discrete latent variables. According to the tutorial on MCMC techniques for HMM inference [36] and other previous works on HMM [37], [38], [39], [40], we can easily develop a point-wise collapsed Gibbs sampling algorithm that iteratively samples these discrete latent variables, i.e., $g_{i,n}$, $z_{i,n}$, $g'_{j,m}$, $z'_{j,m}$. At each iteration for source user n, we first sample the latent mode $z_{i,n}$ at each time slot i from its full conditional posterior distribution $p(z_{i,n}|Rest)$ and then resample each sequential schema indicator $g_{i,n}$ from its full conditional posterior distribution $p(g_{i,n}|Rest)$, as shown in Table 1. A similar process can be conducted for the target user m, which is omitted due to the page limit. After the burn-in period, we can update the transition matrix parameter $\mathbf{A}_g$ and the exponential rate parameter $\mathbf{B}_k$ using the expectation under their posterior distributions, respectively, as follows:

$$\mathbf{A}_{g,k_1,k_2} = \frac{\sum_{n=1}^{N} C_{n,g,k_1,k_2} + d}{\sum_{n=1}^{N} \sum_{k_2=1}^{K} C_{n,g,k_1,k_2} + Kd}, \tag{7}$$

$$\mathbf{B}_k = \frac{c_1 + \sum_{n=1}^{N} \sum_{i}^{I_n} \Delta(z_{i,n} = k)}{c_2 + \sum_{n=1}^{N} \sum_{i}^{I_n} \delta_{i,n}\Delta(z_{i,n} = k)}. \tag{8}$$

3.5.2 Variational Inference

For convenient inference of Poisson factorization (BPF), we first augment our model with auxiliary variables [34]. Specifically, we introduce G latent variables $l_{g,n,m} \sim \text{Poisson}(\boldsymbol{\theta}_{g,n}\boldsymbol{\theta'}_{g,m})$, which are integers such that $y_{n,m} = \sum_g l_{g,n,m}$. Then we can calculate the joint likelihood as follows:

TABLE 2
Latent Variables, Complete Conditionals, and Variational Parameters

Latent variable	Type	Complete conditional	Variational parameters
$\theta_{g,n}$	Gamma	$< a + \pi_{g,n} + l_{g,n,\cdot}, \; b + \theta'_{g,\cdot} >$	$< \widetilde{\theta}^s_{g,n}, \widetilde{\theta}^r_{g,n} >$
$\theta'_{g,m}$	Gamma	$< a' + \pi'_{g,m} + \sum_n l_{g,n,m}, \; b' + \sum_n \theta_{g,n} >$	$< \widetilde{\theta}'^s_{g,m}, \widetilde{\theta}'^r_{g,m} >$
$l_{n,m}$	Multi	$\log \theta_{g,n} + \log \theta'_{g,m}$	$\phi_{n,m}$

$$
\begin{aligned}
& p(\boldsymbol{\theta}, \boldsymbol{\theta}', l, \boldsymbol{\pi}, \boldsymbol{\pi}', g, g^{'} | \boldsymbol{\Theta}, \mathbf{Y}) \\
= & \prod_{g} \left[\prod_{n} p(\theta_{g,n} | \pi_{g,n}, a, b) \prod_{m} p(\theta'_{g,m} | \pi'_{g,m}, a', b') \right. \\
& \left. \prod_{n,m} p(l_{g,n,m} | y_{n,m}, \theta_{g,n}, \theta'_{g,m}) \right] \prod_{n} \left[p(\pi_n | \alpha) \prod_{i} p(g_{i,n} | \pi_n) \right] \\
& \prod_{m} \left[p(\pi'_m | \alpha') \prod_{j} p(g'_{j,m} | \pi'_m) \right].
\end{aligned} \tag{9}
$$

Variational Family. We define the mean-field variational family q over the latent variables $\boldsymbol{\theta}, \boldsymbol{\theta}', l$ to approximate the true posterior of latent variables in BPF as follows:

$$
q(\theta, \theta^{'}, l) = \prod_{g,n} q(\theta_{g,n}) \prod_{g,m} q(\theta^{'}_{g,m}) \prod_{n,m} q(l_{n,m}). \tag{10}
$$

Due to the conditionally conjugate characteristics of $\theta_{g,n}, \theta^{'}_{g,m}$, the variational factors for $\theta_{g,n}, \theta^{'}_{g,m}$ are all Gamma distributions—the same as their complete conditional distributions—with freely set shape and rate variational parameters:

$$
q(\theta_{g,n}) = \mathrm{Gamma}(\widetilde{\theta}^{(s)}_{g,n}, \widetilde{\theta}^{(r)}_{g,n}), \tag{11}
$$

$$
q(\theta'_{g,m}) = \mathrm{Gamma}(\widetilde{\theta}'^{(s)}_{g,m}, \widetilde{\theta}'^{(r)}_{g,m}). \tag{12}
$$

We denote the shape parameter of Gamma distribution as $\theta^{(s)}_{\cdot,\cdot}$ and the rate parameter as $\theta^{(r)}_{\cdot,\cdot}$.

For the auxiliary Poisson variables, $l_{n,m}$ is a G-dimensional latent vector of Poisson counts, which is governed by a multinomial distribution when conditioned on their observed sum $y_{n,m}$ [41], [42]. Thus, the variational factor for $l_{n,m}$ is a Multinomial $\mathrm{Mult}(y_{n,m}, \boldsymbol{\phi}_{n,m})$ where the variational parameter $\boldsymbol{\phi}_{n,m}$ is a point on the G-simplex:

$$
q(l_{n,m}) = \mathrm{Multi}(y_{n,m}, \boldsymbol{\phi}_{n,m}). \tag{13}
$$

We show the complete conditionals and its corresponding variational parameters of $\boldsymbol{\theta}_{g,n}$, $\boldsymbol{\theta}'_{g,m}$, $l_{n,m}$ in Table 2.

3.5.3 Nonlinear Optimization

Given the Dirichlet-Multinomial-Gamma distributions, i.e., $p(\boldsymbol{\pi}_n | \alpha)$, $p(g_{i,n} | \boldsymbol{\pi}_n)$ and $p(\boldsymbol{\theta}_n | \boldsymbol{\pi}_n, a, b)$, which all involve the estimation of the latent schema representation $\boldsymbol{\pi}_n$, we use nonlinear optimization technique [43] to optimize the non-conjugate parameter $\boldsymbol{\pi}_n$.

We can easily obtain the complete log likelihood $\mathcal{L}$ using Eq. (9) as follows:

$$
\begin{aligned}
& \mathcal{L}(\pi, \pi') \\
= & \sum_{g,n} \log p(\theta_{g,n} | \pi_{g,n}, a, b) + \sum_{g,m} \log p(\theta'_{g,m} | \pi'_{g,m}, a', b') \\
& + \sum_{n} \left[\log p(\pi_n | \alpha) + \sum_{i} \log p(g_{i,n} | \pi_n) \right] \\
& + \sum_{m} \left[\log p(\pi'_m | \alpha') + \sum_{j} \log p(g'_{j,m} | \pi'_m) \right] + Const \\
= & \sum_{n=1}^{N} \sum_{g=1}^{G} \left\{ (\alpha_g - 1 + C_{g,n}) \log \pi_{g,n} + \pi_{g,n} \log b \right. \\
& \left. + \pi_{g,n} \log \theta_{g,n} - \log \Gamma(\pi_{g,n} + a) \right\} + \sum_{m=1}^{M} \sum_{g=1}^{G} \left\{ (\alpha'_g \right. \\
& - 1 + C_{g,m}) \log \pi'_{g,m} + \pi'_{g,m} \log b' + \pi'_{g,m} \log b' \\
& \left. + \pi'_{g,m} \log \theta'_{g,m} - \log \Gamma(\pi'_{g,m} + a') \right\} + Const,
\end{aligned} \tag{14}
$$

where *Const* is a constant variable that does not include these random variables, $C_{g,n}$ is the count of g appearing in source n's sequence, $C_{g',m}$ is the count of g' appearing in target m's sequence.

Since the closed updating formula for the non-conjugate variable π_n does not exist, we use projection gradient method to approximate the parameters, with the gradients calculated as follows:

$$
\begin{aligned}
\frac{\partial \mathcal{L}_{\pi_{g,n}}}{\partial \pi_{g,n}} &= \frac{\alpha_g + C_{g,n} - 1}{\pi_{g,n}} + \log b + \log \theta_{g,n} - \Psi(a + \pi_{g,n}), \\
\frac{\partial \mathcal{L}_{\pi'_{g,m}}}{\partial \pi'_{g,m}} &= \frac{\alpha'_g + C_{g,m} - 1}{\pi'_{g,m}} + \log b' + \log \theta'_{g,m} - \Psi(a' + \pi'_{g,m}),
\end{aligned} \tag{15}
$$

where we can use the expectation of variational parameters $\theta_{g,n} \approx \frac{\widetilde{\theta}^s_{g,n}}{\widetilde{\theta}^r_{g,n}}$, $\theta'_{g,m} \approx \frac{\widetilde{\theta}'^s_{g,m}}{\widetilde{\theta}'^r_{g,m}}$, and $\Psi(\cdot)$ denotes the first derivative of the function $\log \Gamma(\cdot)$.

Algorithm 1. Hybrid Gibbs-Variational Algorithm

```
Initialize $\{g^{(0)}_n\}$, $\{g'^{(0)}_m\}$, $\{z^{(0)}_n\}$, $\{z'^{(0)}_m\}$ using 100 iterations gibbs sampling described in Table 1.
Initialize $\pi^{(0)}_n$ and $\pi'^{(0)}_m$ using Eqs. (16) and (17), respectively.
for each iteration $it \in [1, maxIter]$ do
    Resample $\{g^{(it)}_n\}$, $\{g'^{(it)}_m\}$, $\{z^{(it)}_n\}$, $\{z'^{(it)}_m\}$ in a similar fashion as in Table 1.
    Update variational parameters $\widetilde{\theta}^{(it)}_{g,n}$, $\widetilde{\theta}'^{(it)}_{g,m}$, $\phi_{n,m}$ as in Table 2.
    Update latent user representation $\pi^{(it)}_n$, $\pi'^{(it)}_m$ using projection gradient as described in Section 3.5.3.
end for
Update $\mathbf{A}$, $\mathbf{B}$ using Eqs. (7) and (8), respectively.
```

The details of the hybrid Gibbs-Variational estimation method are shown in Algorithm 1. Note that, we use the Bayesian estimations from Gibbs Sampling as the initializations of latent representations $\boldsymbol{\pi}_n$, $\boldsymbol{\pi'}_m$ for source and target users, respectively, which in turn serve as priors for variational inference:

$$\boldsymbol{\pi}_{g,n}^{(0)} = \frac{C_{g,n} + \alpha_g}{C_n + G\alpha_g}, \tag{16}$$

$$\boldsymbol{\pi}'^{(0)}_{g,m} = \frac{C_{g,m} + \alpha'_g}{C_m + G\alpha'_g}. \tag{17}$$

3.6 Fraud Detection Criterion

Given an interaction network consisting of N sources and M targets, our objective is to detect source users whose behaviors deviate from the normal in terms of sequential and interaction schemas. As the proposed model COSIN suggests, both the sequences and interaction behavior data can be generated to reveal several different behavioral patterns. We generally assume that the proportion of fraudsters is small, thus when the behaviors that do not fit in these patterns, or in other words, when the likelihood of the behavioral data is low, they can be identified as fraudulence.

Therefore, one intuitive choice to evaluate the deviation of behaviors from the constructed model is to compute the marginal data likelihood of each source. In addition, as discussed previously, the time-variate sequential schema can decide the transitional patterns that govern the transitions between latent modes. To account for the distinct transitional patterns, we incorporate the latent sequential schema g_n of each user, which gives the complete log likelihood as follows:

$$\begin{aligned}&\log p(\delta_n, g_n, y_{n,\cdot}|\Theta)\\ =&\log p(\delta_n|\Theta) + \log p(g_n|\Theta) + \log p(y_{n,\cdot}|\Theta).\end{aligned} \tag{18}$$

In this way, we incorporate both sequence and interaction behaviors for fraud detection in the interaction network. In practice, we define the negative complete log likelihood as the anomaly score COSIN_n for each source user n as follows:

$$\mathrm{COSIN}_n = -\log p(\delta_n, g_n, y_{n,\cdot}|\Theta). \tag{19}$$

Then, a source user with higher anomaly score is more likely to be detected as a fraudster.

4 Experiment

In this section we first introduce the baseline methods and then present extensive experimental results on both synthetic and real-world telecommunication data sets. Area under the ROC-Curve (AUC) is used as the evaluation metric for comparative study whenever data label is available. All the experiments are carried out on a Linux Server with 2 Intel Xeon E5-2640 v2 2.00GHz CPUs and 173 GB memory.

4.1 Baseline Methods

Among various fraud detection methods, GLAD [32] is similar to our proposed model by combining both sequential and interaction behaviors into account. We also compare our model with baseline methods that detect fraudulence in singe perspective, either *interaction-based* or *sequence-based*. Moreover, we employ a telecom fraud detection method based on LDA as another baseline method. Several submodels of COSIN are also used as baseline methods to show the effectiveness of the synergistic effect of the full model.

4.1.1 Group Latent Anomaly Detection (GLAD)

GLAD [32] is employed to take both pairwise (i.e., interactions) and point-wise data (i.e., discrete calling inter-arrival time) as input for group anomaly detection. The anomalous score of a group g can be computed as $\mathrm{Score}(g) = -\sum_{n\in g}\mathbb{E}_q$ $[\log p(R_n|\Theta)]$. Further, to align the model with the problem of detecting the individual fraudster, we assign each user n to its most probable group according to the group distribution, $g_* = \mathrm{argmax}_g\pi_{g,n}$. Then the anomalous score of group g_*, i.e., $\mathrm{Score}(g_*)$ can be used to represent the anomalous score of user n.

4.1.2 Interaction-Based Methods

Interaction-based methods aim to detect frauds from the interaction network. We treat the target users that source user n interacts with as feature vector, and represent each source n by an M-dimensional vector $(y_{n1},\ldots,y_{nM})^\top$ in a "bag-of-words" fashion, where y_{nm} is the count that n interacts with m. Given the feature vector of each user, we can implement the following anomaly detection methods.

Robust covariance (RC) [44] assumes that normal data come from a known distribution (i.e., Gaussian distribution) and fits a robust covariance estimate to the data. By defining the shape of normal data, RC gives high anomalous scores to observations which stand far enough from the fitted shape.

One-class support vector machines (OCSVM) [45] extends traditional SVM to adapt for anomaly detection. We use RBF kernel and the normalized signed distances to the separating hyperplane as anomalous scores of data points.

Isolation forest (iForest) [46] detects anomalies by isolating instances with tree structures. The average path length on the tree structures defines the anomalous scores of instances.

Local Outlier Factor (LOF) [47] is a density-based outlier detection algorithm. We use the computed LOF score to represent the abnormality of the observations.

Poisson factorization (PF) is a submodel of our proposed method. This method takes only the interaction data as input and the anomaly score for each source n is defined using negative data log likelihood of observed interaction counts as follows: $S_n = -\log p(y_{n,\cdot}|\Theta)$.

4.1.3 Sequence-Based Methods

In modeling the sequences, we apply HMM and its variates as baseline methods. Note that we only model the sequences of the source users in these methods.

Global hidden Markov model (GlobalHMM) assumes that all users share one global transition matrix and K modes. We define the anomaly score of the source n as: $S_n = -\log p(\delta_n|\Theta)$.

Mixture Transition Hidden Markov model (MTHMM) is indeed a submodel of our proposed method that only models the sequences of source users. It assumes that all the source users share G schemas and K modes. We define the anomaly score of the source n as: $S_n = -\log p(\boldsymbol{\delta}_n, g_n|\Theta)$.

4.1.4 Latent Dirichlet Allocation (LDA) for Telecom Frauds

As proposed by [26], Latent Dirichlet Allocation is employed to build caller's profile signatures in telecom and unexpected deviations from the normal are regarded as telecom frauds. To adapt LDA in our settings, each discrete calling inter-arrival time is treated as one word, each source user can be viewed as one document, and thus the calling inter-arrival time data of all source users can make up a corpus. For training, we fit LDA model on the whole corpus and then use the inferred perplexity of each document (source user) as its anomaly score.

4.1.5 Submodels of COSIN

We introduce two submodels of our proposed COSIN model, i.e., COSIN-Separate and COSIN-Caller.

COSIN-Separate learns sequential schema and interaction schema separately in a similar manner as COSIN in Eq. (19), but without using $\boldsymbol{\pi}$ to align two schemas. For fraud detection criterion, we obtain the anomaly scores for both sequences and interactions, and then add the two scores into a score.

COSIN-Caller and COSIN share the same scoring technique as in Eq. (19), but without modeling target users' sequential schema.

4.2 Fraud Detection on Synthetic Dataset

We generate a synthetic dataset by following the generative process of a suitably defined COSIN model. Specifically, we generate $N = 1000$ sources and $M = 1000$ targets, with 100 (10 percent) anomalous sources injected as fraudsters, which results in a total 1,067,404 transaction records. The sequence and interaction data of both normal and anomalous instances are generated according to the generative process of COSIN with varied parameter settings. Since the generation of targets could be conducted in a similar manner as that of source users, we only detail the simulation process of source users as follows. The number of latent schemas and modes were set as $G = 10$ and $K = 5$, respectively. We defined the anomalous sources as instances that have abnormal sequences and interaction structure. Specifically, we first consider the generation of shared latent representation $\boldsymbol{\pi}_n$. We set the symmetric Dirichlet prior $\alpha = 100$ for generating normal sources' shared latent representation, and set that of anomalous sources to $\alpha = 0.01$. We find this settings can produce latent representations that distinguish the normal from anomalies, i.e., each normal source has multiple different sequence and interaction schemas while all anomalous sources share only one main schema. Moreover, inspired by the motivation that fraudulent users can differ tremendously from normal users in the transitional matrix, we generate $G = 10$ different transitional matrices with the Dirichlet prior $d = 0.01$ for normal sources. While for fraudulent sources, we manually employ a diagonal transitional matrix to represent its sequential schema.

TABLE 3
Performances on Synthetic Dataset

Method	AUC	Stat. Test COSIN vs.
OCSVM	0.6707 ± 0.0000	<0.001
iForest	0.8255 ± 0.0325	<0.001
LOF	0.8014 ± 0.0000	<0.001
RC	0.8615 ± 0.0136	<0.001
PF	0.7551 ± 0.0350	<0.001
GlobalHMM	0.6021 ± 0.0001	<0.001
MTHMM	0.6378 ± 0.0325	<0.001
GLAD	0.5060 ± 0.0180	<0.001
LDA	0.3154 ± 0.0015	<0.001
COSIN-Separate	0.6990 ± 0.0199	<0.001
COSIN-Caller	0.7634 ± 0.0149	<0.001
COSIN	**0.9160 ± 0.0249**	

Parameter Settings for Comparison. For probabilistic generative models, including COSIN, COSIN-Separate, COSIN-Caller, MTHMM, GlobalHMM, PF and GLAD,[1] we set the parameters the same as the parameters that generate this synthetic dataset, i.e., $G = 10$ and $K = 5$, in the training phase. For other baseline methods, i.e., LOF, iForest, OCSVM and RC, we conduct exhaustive grid search to find the optimal parameters for achieving the best fraud detection performance in terms of AUC values. We detail the performances of these four methods under different parameters in the supplemental material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TKDE.2019.2912817.

Performance on Synthetic Dataset. In Table 3, we show fraud detection performances evaluated through average AUC values along with standard deviations over 10 rounds. It can be seen that most interaction-based methods (LOF, iForest, PF) outperform sequence-based methods (MTHMM and GlobalHMM), which shows that the fraudulent patterns in view of interactions are easier to be detected than that of sequences under the synthetic settings. Our proposed COSIN model, which fuses both sequence and interaction data, outperforms all the baseline methods. Moreover, we conduct T-test to show that the outperformances made by COSIN over all baseline methods are statistically significant (p-value < 0.001).

4.3 Fraud Detection on Real-World Telecom Datasets

We present the experiments of our proposed COSIN model on the real-world telecommunication data sets, which are obtained from one of the telecom operators in China. The datasets contain Call Detail Records (CDR) in telecommunication network, with each CDR consisting of transactional details of every individual call, including encrypted phone numbers of both source and target parties, the start timestamp of the call, etc. In real-world data, we find that the telecom network is rather sparse, even a small number of callers can invoke large sum of call events toward distinct targets. Therefore, in order to uncover the behavioral

[1] Note that, we set the number of *groups* in GLAD the same as the number of latent schemas G in COSIN, and the number of *roles* the same as the number of latent modes K.

TABLE 4
Statistics of Real-World Telecom Datasets

	Promotional	Recording	Genuine	Partially Labelled	Large Inter. Network
# Fraudulent callers	93	106	127	197	16,722
# Normal callers	335	877	1,108	9,803	91,940
# Total callers	428	983	1,235	10,000	108,662
# Total CDRs	1,063,561	1,256,745	318,262	3,154,150	4,069,727
Anomaly ratio	0.2173	0.1078	0.1028	0.0197	0.1539

patterns of the callers in telecom network, we first aggregate the target phone numbers by their innate characteristic, i.e., called area codes, with the help from the operator, and then the set of target callees is reduced to a relatively stable set.

4.3.1 Data Description

From Sep. 2017 to Nov. 2018, we randomly collect 500 thousand users who ever invoked at least three calls on real-world telecom records. Then, we extract their full-length calling sequences to form a base dataset consisting of more than 10 million CDRs. Since there exists no ground-truth label for telecom frauds, we resort to the crowdsourcing service provided by the telecom operators, in which ordinary telecom callees can help to tag the calls they have just received as different types of frauds or normal. Based on the collected data, we construct three groups of datasets: i) small but completely-labeled datasets with different types of telecom frauds; ii) a medium-size dataset with only a small proportion of data labeled as frauds; iii) a large-scale completely-labeled dataset to evaluate the effectiveness of COSIN in handling large interaction network. The statistics of all the real-world datasets are presented in Table 4, with detailed characteristics given as follows.

Completely-Labeled Datasets with Different Types of Anomalies. In the crowdsourcing service, the receiver of a call can mark it as a fraudulence with a specific anomalous type, including *Promotional*, *Recording* and *Genuine*, described as follows:

1) *Promotional* callers constantly invoke large quantities of calls in a short time period to targets for promotional purposes, i.e., promoting and selling products or services.
2) *Recording* callers first record their voice as pre-designed scripts and replay their scripts when invoking calls to targets for intentional deception resulting in economic losses of victims.
3) *Genuine* callers are similar to recording callers in terms of their deceptive purposes, but differ in that they do not use pre-designed voice scripts and speak directly to the victims. They usually pretend to be officials working in public sectors such as police, tax bureau, etc., or acquaintances of the callee.

We randomly sample and obtain a completely-labeled dataset with 93 *promotional* callers, 106 *recording* callers, 127 *genuine* callers, 2,320 normal callers and their corresponding calling records. Considering the fact that different fraudulent types have their unique characters, we further split this dataset into three segments with each segment having only one type of fraudsters, and evaluate the detection performance separately.

Partially Labeled Dataset. We randomly sample another 197 fraudulent callers without considering the fraudulent types, and inject them into a medium-size set of unlabeled callers. Finally, we obtain 10,000 callers with 3,154,150 calling records, resulting in a partially-labeled dataset. In this dataset, the labeled fraudulent callers only account for 1 percent of all the users, which can validate the effectiveness the model when only scarce labels are available.

Large Interaction Network Dataset. To study the effectiveness of handling large interaction network, we randomly sample 108,662 callers, among which 16,772 callers are labeled as anomalies, without distinguishing the fraudulent types. The total number of corresponding CDRs reaches 4,069,727.

4.3.2 Fraud Detection Performances

For all the real-world datasets, we show the mean AUC values and standard deviations of all methods in 10 rounds in Table 5. Also, we conduct T-test between COSIN and each baseline for statistical significance testing on each dataset.

Parameter Settings. As a probabilistic model, the parameters of COSIN such as the number of latent schema G and number of latent modes K need to be initially specified but remain unknown in prior. Thus, we use a model scoring method, i.e., Akaike Information Criterion (AIC), to automatically determine the optimal values. The definition of AIC score is given by:

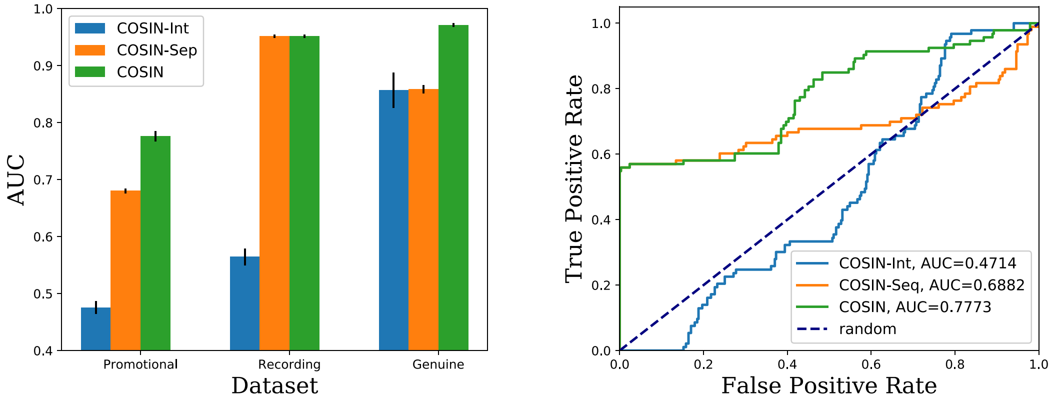
$$AIC(\Delta, Y; \Theta) = -2 \log p(\Delta, Y|\Theta) + 2|\Theta|, \tag{20}$$

where $|\Theta|$ denote the number of free parameters in the model. We can see that AIC balances between the goodness-of-fit and model over-fitting. It rewards goodness-of-fit (as assessed by the likelihood function), but also includes a penalty that is an increasing function of the number of estimated parameters. Theoretically, we choose G and K values with a minimum AIC value by performing a two-dimensional grid search. For baselines of probabilistic generative models, i.e., COSIN-Separate, COSIN-Caller, MTHMM, GlobalHMM, PF and GLAD, we take a similar model scoring method using AIC value to search for their best parameters. Empirically, we find that on *Promotional* dataset for all probabilistic generative methods, $G = 5$ and $K = 10$ not only achieve reasonable AIC values but also show acceptable efficiency. Therefore, we set $G = 5$ and $K = 10$ as the default settings for all the probabilistic generative methods. While for the four interaction-based methods, i.e., OCSVM, iForest, LOF and RC, we take a similar exhaustive search approach as described in synthetic experiments to choose the best parameters.

TABLE 5
Fraud Detection Performances on Real-World Telecom Datasets

Method	Promotional		Recording		Genuine		Partially Labelled		Large Interaction Network	
	AUC	Stat. Test COSIN vs.	AUC	Stat. Test COSIN vs.	AUC	Stat. Test COSIN vs.	AUC	Stat. Test COSIN vs.	AUC	Stat. Test COSIN vs.
OCSVM	0.6945 ± 0.0000	<0.001	0.8437 ± 0.0000	<0.001	0.5559 ± 0.0000	<0.001	0.7894 ± 0.0000	<0.001	0.4875 ± 0.0000	<0.001
iForest	$\underline{0.7195} \pm \underline{0.0208}$	<0.001	0.8255 ± 0.0100	<0.001	0.7930 ± 0.0373	<0.001	0.7836 ± 0.0095	<0.001	0.3882 ± 0.0312	<0.001
LOF	0.5312 ± 0.0000	<0.001	0.4464 ± 0.0000	<0.001	0.8944 ± 0.0000	<0.001	0.3470 ± 0.0000	<0.001	0.6857 ± 0.0000	<0.001
RC	0.6681 ± 0.0528	<0.001	0.8408 ± 0.0204	<0.001	0.8922 ± 0.0308	<0.001	0.8380 ± 0.0388	<0.05	N/A	N/A
PF	0.4360 ± 0.0269	<0.001	0.5467 ± 0.0404	<0.001	0.7920 ± 0.0227	<0.001	0.5304 ± 0.0411	<0.001	0.6549 ± 0.0158	<0.001
GlobalHMM	0.6521 ± 0.0015	<0.001	0.9041 ± 0.0002	<0.001	0.7089 ± 0.0005	<0.001	0.8487 ± 0.0001	<0.001	0.7121 ± 0.0023	<0.001
MTHMM	0.6791 ± 0.0023	<0.001	$\underline{0.9330} \pm \underline{0.0049}$	<0.001	0.8565 ± 0.0071	<0.001	$\underline{0.8635} \pm \underline{0.0015}$	<0.001	$\underline{0.7614} \pm \underline{0.0003}$	<0.001
GLAD	0.6793 ± 0.0347	<0.05	0.8077 ± 0.0345	<0.001	0.5719 ± 0.1119	<0.001	0.7573 ± 0.0577	<0.001	N/A	N/A
LDA	0.3705 ± 0.0112	<0.001	0.0650 ± 0.0016	<0.001	0.3584 ± 0.0092	<0.001	0.1247 ± 0.0062	<0.001	0.1492 ± 0.0166	<0.001
COSIN-Separate	0.6035 ± 0.0492	<0.001	0.7607 ± 0.0096	<0.001	$\underline{0.9587} \pm \underline{0.0087}$	<0.001	0.7286 ± 0.0108	<0.001	0.7519 ± 0.0037	<0.001
COSIN-Caller	0.6565 ± 0.0223	<0.001	0.7715 ± 0.0083	<0.001	0.9580 ± 0.0087	<0.001	0.7470 ± 0.0150	<0.001	0.7504 ± 0.0006	<0.001
COSIN	**0.7760** ± **0.0090**		**0.9518** ± **0.0031**		**0.9711** ± **0.0032**		**0.8689** ± **0.0026**		**0.7659** ± **0.0008**	

***N/A**: failure due to the out-of-memory error. The best results are in bold and the second best underlined. The average AUC values along with their standard deviations in 10 rounds are shown. The p-values by T-test for COSIN versus each baseline have also been given.*



(a) AUC on different datasets.

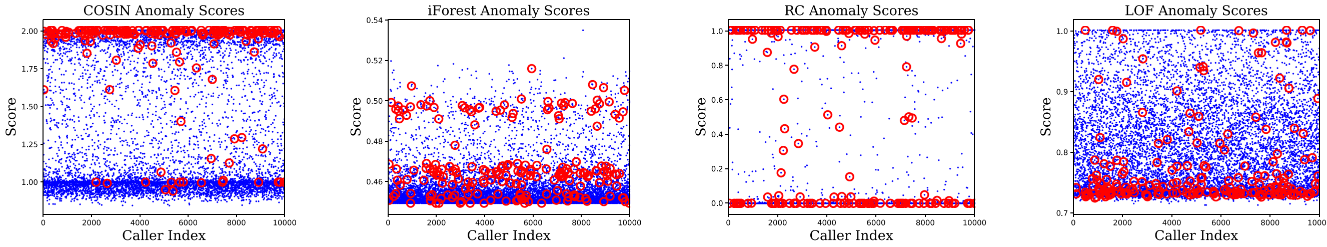
(b) ROC curves on promotional.

Fig. 3. Performance of COSIN using different scoring methods for detecting different types of telecom frauds.

Detection Performances for Different Types of Telecom Frauds. As shown in Table 5, COSIN consistently outperforms all the baselines, which well demonstrates the effectiveness of COSIN in detecting different types of telecom frauds. The advantages of COSIN over other baselines are statistically significant (p-value $<$ 0.001 by T-test) except for the test of COSIN versus GLAD on *Promotional* dataset, which is owing to the tremendous fluctuations of GLAD in multiple rounds of experiments. This demonstrates the robustness of the outperformance of COSIN against baseline methods.

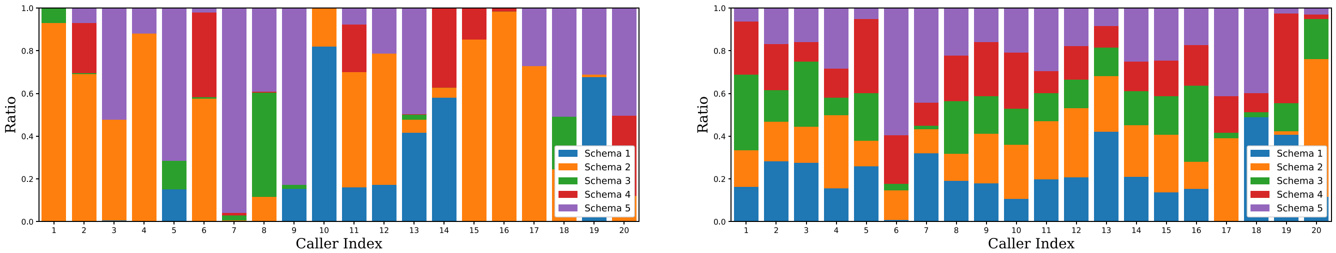
In modeling the sequential behaviors, we find that by introducing a higher level of latent schema, MTHMM achieves better fraud detection performance than the conventional HMM with only one layer of latent states, i.e., GlobalHMM. This is in accordance with our intuition that a higher level of latent schema can capture the time-variate and individual oriented transitional matrix, so as to better represent the sequential schemas and distinguish frauds from the normal. Moreover, it is worth noting that the full model COSIN consistently outperforms the submodel COSIN-Separate. This well demonstrates that the synergistic effect of behavioral sequences and interactions is accurately captured, which in turn enhances the detection performance significantly. Also, simultaneously modeling the sequential behaviors of both source and target users is indeed critically important, which is validated by the consistent advantage of COSIN over the variant version, COSIN-Caller. This is due to the fact that receivers of fraudulent calls can also differ from normal receivers in sequential and interaction behaviors.

Additionally, we further show the robustness of COSIN under different scoring methods by decomposing anomalous scores of COSIN into interactions and sequences scores, respectively. We take the data likelihood of sequential behaviors under COSIN as the anomaly score of COSIN-Seq, while COSIN-Int refers to the score that only employs the data likelihood of the interactions under COSIN . In Fig. 3a, it is interesting to see that COSIN-Seq achieves highest AUC in detecting fraudulent sequential schema on *Recording* dataset, while the fraud detection from the perspective of interaction schema on *Genuine* dataset achieves high value of AUC with the scoring function COSIN-Int. This might be explained by that *recording* callers are active in calling large sum of targets sequentially and are easy to be distinguished from the normal in terms of their calling sequences, whereas the *genuine* callers usually have some



(a) Anomaly scores of COSIN. (b) Anomaly scores of iForest. (c) Anomaly scores of RC. (d) Anomaly scores of LOF.

Fig. 4. Performance on partially-labeled dataset. Red circles denote the 179 labeled anomalous callers among the total 10,000 callers.



(a) Distribution of sequential schema for normal source users. (b) Distribution of sequential schema for fraudulent source users.

Fig. 5. Learned distribution of sequential schema for each normal (a) and anomaly (b). We only present top-20 normal/anomalous sources sorted by the anomalous score. Different colors in the bar chart represent different schemas.

abnormal strategies in choosing victims and the abnormality is thus manifested more significantly in the interaction schema. In Fig. 3b, it is also notable that COSIN-Seq or COSIN-Int performs even worse than a random classifier in some parts of the ROC curve, while in contrast COSIN performs well along the whole curve. This should be owing to the synergistic power of COSIN in collectively modeling different behaviors.

Detection Performances on Partially Labeled Dataset. We can see from Table 5, COSIN achieves the best performance in detecting the labeled fraudsters in terms of the AUC measure. However, since the experimental dataset is a partially labeled set that only accounts for a small proportion, we cannot obtain the accurate AUC of the whole dataset. As such, we further plot the anomalous scores of all the callers output from several competitive methods including our proposed COSIN, iForest, RC and LOF in Fig. 4, where the labeled ground-truth fraudsters are circled in red. In Fig. 4a, most of the ground-truth fraudulent callers lie in the upper space of the figure, while they are more dispersive in other sub-figures for other methods. This implies that COSIN can better distinguish fraudulent behaviors from unlabeled callers than other competitive baselines.

Detection Performances in Large Interaction Network. Table 5 shows that COSIN also achieves the best performances in the large-scale dynamic interaction network with statistical significance (p-value<0.001), which scales up to 100 thousand callers. In contrast, some baseline methods, e.g., GLAD and RC, fail to deliver results on this dataset due to the out-of-memory error. In supplemental material, we further conduct experiments on synthetic datasets to demonstrate the efficiency of our proposed method COSIN in handling large interaction network. The results show that COSIN can scale up to more than 10,000,000 sources/targets in our experimental environment, and the training for a network with 100,000 nodes can be completed within 15 minutes.

4.3.3 Case Study

We finally conduct a case study on the *Recording* dataset with $G = 5$ and $K = 10$ in order to show the interpretability of COSIN. By inferring the parameters from COSIN model, the learned distribution of the sequential schema assignments for each source user is shown in Fig. 5. We can see that the learned sequential schema distributions of the normal sources differ significantly from those of the fraudulent sources, i.e., normal sources usually have one main sequential schema whereas fraudulent sources have more evenly distributed latent schemas.

Fig. 6 shows distinct mode transition patterns with respect to different sequential schemas, representing the probability of current mode i transiting to the mode j at next time slot under a specific sequential schema. For instance, under sequential schema 3, entry (6, 1), entry (10, 1), entry (7, 7) take most of the probability mass, indicating that there is a high probability of transiting from mode 6, 10 to mode 1 or remaining the same at mode 7. We can also see much probability mass lies in the uppper left diagonal of the transitional matrix of schema 3; while in contrast, mode transitions in sequential schema 5 concentrate along the bottom right diagonal of the matrix

In addition, we can further interpret particular calling modes semantically. Specifically, we group the time intervals between every two consecutive calls of all the callers that are assigned with the same latent calling mode z. For each calling mode, we plot the probability density of the observed time intervals in Fig. 7, from which we can see that some calling modes (e.g., modes 5, 7, 9, 10) contain large and evenly distributed mass in short time intervals while other calling modes (i.e., mode 8) show some periodic

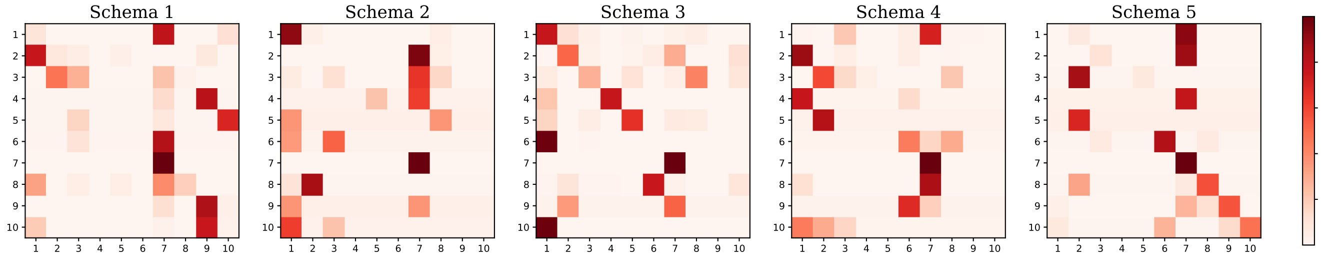


Fig. 6. Learned model transition patterns for each sequential schema. The *y*-axis represents where the mode transits from, and the *x*-axis is the mode it would transit to.

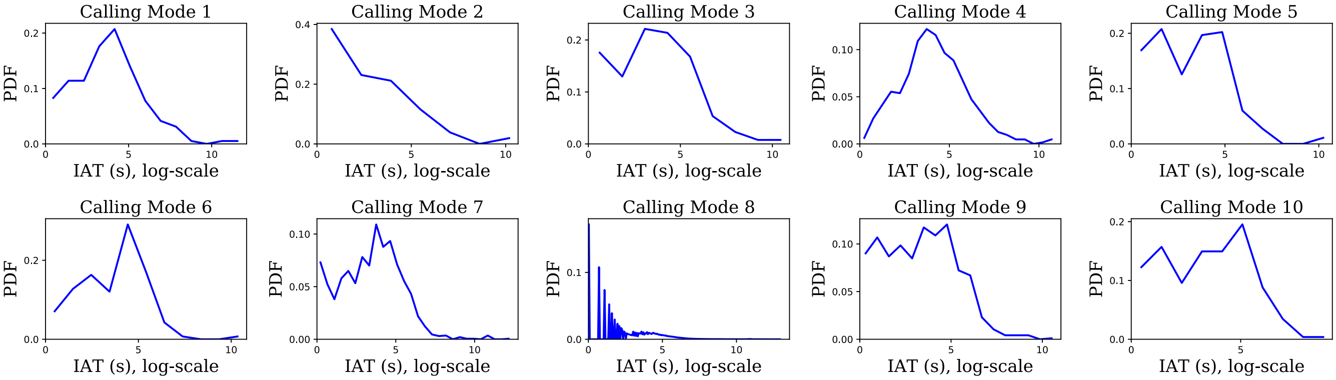


Fig. 7. Distribution of inter-arrival time (IAT) for each calling mode. The *y*-axis represents the probability density, and the *x*-axis represents the logarithmic values of time intervals.

patterns, demonstrating that the proposed COSIN is capable of identifying different sequential behaviors and the corresponding behavioral patterns.

5 Conclusion

Fraud detection is generally a challenging task since fraudsters usually hide their malicious behaviors in the large amount of normal behaviors. However, fraudulence in dynamic interaction network can be manifested if both sequence and interaction behaviors are considered in a synergistic framework. Therefore, in this paper, we investigated the fraud detection problem in dynamic interaction network by modeling the sequences and the interactions in a collective probabilistic model COSIN. On one hand, the sequential behavior of each individual user was generated in a hierarchical HMM manner. We introduced a latent sequential schema to capture the time-variate and individual-oriented transitions of sequential behaviors, and meanwhile a latent mode was introduced to represent the observed values for the behaviors. On the other hand, the interaction network between pairwise source and target users was modeled by a Poisson factorization, with the interaction schema shifted from the sequential schema. A hybrid Gibbs-Variational algorithm was proposed for efficient parameter estimation.

The model was validated on both synthetic and real-world telecom data sets, either labeled or partially labeled. The results showed that the proposed model COSIN outperformed other baseline methods in different experimental settings and was capable of capturing the transitional patterns and different modes. The model is applicable to various fraud detection scenarios such as review spam detection, money laundering detection where two parties interact dynamically as an interaction network. Moreover, we plan to extend the model by incorporating heterogeneous data such as text, pictures to further improve the performance.

Acknowledgments

Guannan Liu's work is supported by the National Science Foundation of China (NSFC) under Grant 71701007. Junjie Wu's work is partially supported by NSFC under Grants 71725002, 71531001, U1636210, 71490723. Hong Li's work is supported by NSFC under Grant 71471009

References

[1] I. Melnyk, A. Banerjee, B. Matthews, and N. Oza, "Semi-markov switching vector autoregressive model-based anomaly detection in aviation systems," in *Proc. KDD*, 2016, pp. 1065–1074.

[2] P. Bonacich and P. Lloyd, "Eigenvector-like measures of centrality for asymmetric relations," *Soc. Netw.*, vol. 23, no. 3, pp. 191–201, 2001.

[3] K. Henderson, B. Gallagher, T. Eliassi-Rad, H. Tong, S. Basu, L. Akoglu, D. Koutra, C. Faloutsos, and L. Li, "Rolx: Structural role extraction & mining in large graphs," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 1231–1239.

[4] L. Akoglu and C. Faloutsos, "Event detection in time series of mobile communication graphs," in *Proc. Army Sci. Conf.*, 2010, pp. 77–79.

[5] R. A. Rossi, B. Gallagher, J. Neville, and K. Henderson, "Modeling dynamic behavior in large evolving graphs," in *Proc. 6th ACM Int. Conf. Web Search Data Mining*, 2013, pp. 667–676.

[6] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection for discrete sequences: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 5, pp. 823–839, May 2012.

[7] S. Budalakoti, A. N. Srivastava, and M. E. Otey, "Anomaly detection and diagnosis algorithms for discrete symbol sequences with applications to airline safety," *Trans. Syst. Man Cybern. Part C*, vol. 39, no. 1, pp. 101–113, Jan. 2009.

[8] V. Chandola, V. Mithal, and V. Kumar, "Comparative evaluation of anomaly detection techniques for sequence data," in *Proc. 8th IEEE Int. Conf. Data Mining*, Dec. 2008, pp. 743–748.

[9] C. Zhang, K. Zhang, Q. Yuan, L. Zhang, T. Hanratty, and J. Han, "Gmove: Group-level mobility modeling using geo-tagged social media," in *Proc. 22nd ACM SIGKDD Int. Conf.*, 2016, pp. 1305–1314.

[10] L. Rabiner and B. Juang, "An introduction to hidden markov models," *IEEE ASSP Mag.*, vol. AM-3, no. 1, pp. 4–16, Jan. 1986.

[11] J. Janssen and N. Limnios, *Semi-Markov Models and Applications*. New York, NY, USA: Springer Science & Business Media, 2013.

[12] B. Perozzi, L. Akoglu, P. Iglesias Sánchez, and E. Müller, "Focused clustering and outlier detection in large attributed graphs," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 1346–1355.

[13] E. Mller, P. I. Snchez, Y. Mlle, and K. Bhm, "Ranking outlier nodes in subspaces of attributed graphs," in *Proc. IEEE 29th Int. Conf. Data Eng. Workshops*, Apr. 2013, pp. 216–222.

[14] X. He, M. Gao, M.-Y. Kan, and D. Wang, "Birank: Towards ranking on bipartite graphs," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 1, pp. 57–71, Jan. 2017.

[15] S. Pandit, D. H. Chau, S. Wang, and C. Faloutsos, "Netprobe: A fast and scalable system for fraud detection in online auction networks," in *Proc. 16th Int. Conf. World Wide Web*, 2007, pp. 201–210.

[16] J. C. Neil, "Scan statistics for the online discovery of locally anomalous subgraphs," Ph.D. dissertation, Albuquerque, NM, USA, 2011, aAI3474569.

[17] X. Liu and X. Wang, "A network embedding based approach for telecommunications fraud detection," in *Proc. Int. Conf. Cooperative Des. Vis. Eng.*, 2018, pp. 229–236.

[18] Y.-C. Chang, K.-T. Lai, S.-C. T. Chou, and M.-S. Chen, "Mining the networks of telecommunication fraud groups using social network analysis," in *Proc. IEEE/ACM Int. Conf. Adv. Soc. Netw. Anal. Mining*, 2017, pp. 1128–1131.

[19] J. Sun, D. Tao, and C. Faloutsos, "Beyond streams and graphs: Dynamic tensor analysis," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2006, pp. 374–383.

[20] D. Koutra, E. E. Papalexakis, and C. Faloutsos, "Tensorsplat: Spotting latent anomalies in time," in *Proc. 16th Panhellenic Conf. Informat.*, 2012, pp. 144–149.

[21] E. E. Papalexakis, C. Faloutsos, and N. D. Sidiropoulos, "Parcube: Sparse parallelizable tensor decompositions," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2012, pp. 521–536.

[22] M. Araujo, S. Günnemann, S. Papadimitriou, C. Faloutsos, P. Basu, A. Swami, E. E. Papalexakis, and D. Koutra, "Discovery of "comet" communities in temporal and labeled graphs com^2," *Knowl. Inf. Syst.*, vol. 46, no. 3, pp. 657–677, Mar. 2016.

[23] S. Ranshous, S. Shen, D. Koutra, S. Harenberg, C. Faloutsos, and N. F. Samatova, "Anomaly detection in dynamic networks: A survey," *WIREs Comput. Statist.*, vol. 7, no. 3, pp. 223–247, May 2015.

[24] J. Sun, C. Faloutsos, C. Faloutsos, S. Papadimitriou, and P. S. Yu, "Graphscope: Parameter-free mining of large time-evolving graphs," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2007, pp. 687–696.

[25] H. Tong, S. Papadimitriou, J. Sun, P. S. Yu, and C. Faloutsos, "Colibri: Fast mining of large static and dynamic graphs," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 686–694.

[26] D. Xing and M. Girolami, "Employing latent dirichlet allocation for fraud detection in telecommunications," *Pattern Recognit. Lett.*, vol. 28, no. 13, pp. 1727–1734, Oct. 2007.

[27] D. Olszewski, "Employing kullback-leibler divergence and latent dirichlet allocation for fraud detection in telecommunications," *Intell. Data Anal.*, vol. 16, no. 3, pp. 467–485, May 2012.

[28] L. Xiong, B. Póczos, J. G. Schneider, A. Connolly, and J. VanderPlas, "Hierarchical probabilistic models for group anomaly detection," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2011, pp. 789–797.

[29] L. Xiong, B. Póczos, and J. Schneider, "Group anomaly detection using flexible genre models," in *Proc. 24th Int. Conf. Neural Inf. Process. Syst.*, 2011, pp. 1071–1079.

[30] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[31] H. Soleimani and D. J. Miller, "Atd: Anomalous topic discovery in high dimensional discrete data," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 9, pp. 2267–2280, Sep. 2016.

[32] R. Yu, X. He, and Y. Liu, "Glad: Group anomaly detection in social media analysis," *ACM Trans. Knowl. Discovery Data*, vol. 10, no. 2, pp. 18:1–18:22, Oct. 2015.

[33] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[34] P. Gopalan, J. M. Hofman, and D. M. Blei, "Scalable recommendation with hierarchical poisson factorization," in *Proc. 31st Conf. Uncertainty Artif. Intell.*, 2015, pp. 326–335.

[35] P. Wang and P. Blunsom, "Collapsed variational Bayesian inference for hidden Markov models," presented at the *AISTATS*, Scottsdale, AZ, USA, 2013.

[36] J. Besag, *An Introduction to Markov Chain Monte Carlo Methods*. New York, NY, USA: Springer, 2004, pp. 247–270.

[37] J. Gao and M. Johnson, "A comparison of bayesian estimators for unsupervised hidden markov model pos taggers," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2008, pp. 344–352.

[38] S. Goldwater and T. Griffiths, "A fully bayesian approach to unsupervised part-of-speech tagging," in *Proc. 45th Annu. Meet. Assoc. Comput. Linguistics*, Jun. 2007, pp. 744–751.

[39] M. Johnson, "Why doesn't em find good hmm pos-taggers?" in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn.*, Jun. 2007, pp. 296–305.

[40] M. Johnson, T. Griffiths, and S. Goldwater, "Bayesian inference for PCFGs via Markov chain Monte Carlo," in *Proc. Human Lang. Technol. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, Apr. 2007, pp. 139–146.

[41] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Intell. Neuroscience*, vol. 2009, pp. 4:1–4:17, Jan. 2009.

[42] N. L. Johnson, A. W. Kemp, and S. Kotz, *Univariate Discrete Distributions*, vol. 44. Hoboken, NJ, USA: Wiley, 2005.

[43] D. Bertsekas, *Nonlinear Programming*. Belmont, MA, USA: Athena Scientific, 1999.

[44] P. J. Rousseeuw and K. V. Driessen, "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, vol. 41, no. 3, pp. 212–223, Aug. 1999.

[45] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, Jul. 2001.

[46] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation-based anomaly detection," *ACM Trans. Knowl. Discovery Data*, vol. 6, no. 1, pp. 3:1–3:39, Mar. 2012.

[47] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: Identifying density-based local outliers," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2000, pp. 93–104.

Hao Lin received the bachelor's degree in management information systems from Beihang University, in 2013. He is currently working toward the PhD degree in the School of Economics and Management, Beihang University, China. His research interests include data mining and machine learning, with special interests in user modeling and heterogeneous data fusion.

Guannan Liu received the PhD degree from Tsinghua University, China. He is currently an assistant professor with the Department of Information Systems, Beihang University, Beijing, China. His research interests include data mining, social networks, and business intelligence. His work has been published in the *IEEE Transactions on Knowledge and Data Engineering*, the *ACM Transactions on Knowledge Discovery from Data*, the *ACM Transactions on Intelligent Systems and Technology*, *Decision Support Systems*, *Neurocomputing*, etc, and also in the conference proceedings such as KDD, ICDM, SDM, etc.

Junjie Wu received the PhD degree in management science and engineering from Tsinghua University. He is currently a full professor with the Information Systems Department, Beihang University, and the director of the Research Center for Data Intelligence (DIG). His general area of research is data mining and complex networks. He is the recipient of the NSFC Distinguished Young Scholars award and the MOE Changjiang Young Scholars award in China.

Yuan Zuo received the PhD degree from Beihang University, Beijing, China, in 2017. He is currently a post-doctor with the Information Systems Department, Beihang University. His research interests include topic modeling and social computing.

Xin Wan received the BE and PhD degrees in electronic science and technology from the Department of Electronic Engineering, Tsinghua University, Beijing, in 2007 and 2013, respectively. He is currently a senior engineer of the National Computer Network Emergency Response Technical Team / Coordination Center of China (CNCERT/CC). His research interests include network optimization, pattern recognition, and data mining in telecom networks.

Hong Li received the PhD degree from Jilin University, Jilin, China. She is currently an associate professor with the Information Systems Department, Beihang University. Her research interests include data mining and social computing.