ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED
SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 2, 2024

Composite Behavioral Modeling for Identity Theft Detection in Online Social Networks

[1] P. SRINIVASA REDDY, [2] GUDDATI S V S N GANESH

[1](Assistant Professor), MCA, S.V.K.P & Dr K.S. Raju Arts & Science College

[2]MCA, scholar, S.V.K.P & Dr K.S. Raju Arts & Science College

ABSTRACT

In this work, we aim at building a bridge from coarse behavioral data to an effective, quick-response, and robust behavioral model for online identity theft detection. We concentrate on this issue in online social networks (OSNs) where users usually have composite behavioral records, consisting of multidimensional low-quality data, e.g., offline check-ins and online user-generated content (UGC). As an insightful result, we validate that there is a complementary effect among different dimensions of records for modeling users' behavioral patterns. To deeply exploit such a complementary effect, we propose a *joint* (instead of *fused*) model to capture both online and offline features of a user's composite behavior. We evaluate the proposed joint model by comparing it with typical models and their fused model on two real-world datasets: Foursquare and Yelp. The experimental results show that our model outperforms the existing ones, with the area under the receiver operating characteristic curve (AUC) values 0.956 in Foursquare and 0.947 in Yelp, respectively. Particularly, the *recall* (true positive rate) can reach up to 65.3% in Foursquare and 72.2% in Yelp with the corresponding *disturbance rate* (false-positive rate) below 1%. It is worth mentioning that these performances can be achieved by examining only one composite behavior, which guarantees the low response latency of our method. This study would give the cybersecurity community new insights into whether and how real-time online identity authentication can be improved via modeling users' composite behavioral patterns.

1.INTRODUCTION

WITH the rapid development of the Internet, more and more affairs, e.g., mailing, health caring , shopping , booking hotels, and purchasing tickets, are handled online.Meanwhile, the Internet also brings sundry potential risks of invasions, such as losing financial information , identity theft, and privacy leakage . Online accounts serve as the agents of users in the cyber world. Online identity theft is a typical online crime which is the deliberate use of another person's account, usually as a method to gain a financial advantage or obtain credit and other benefits in another person's name. As a matter of fact, compromised accounts are usually the portals of most cybercrimes [1], such as blackmail, fraud, and spam . Thus, identity theft detection is essential to guarantee users' security in the cyber world.

Traditional identity authentication methods are mostly based on access control schemes, e.g., passwords and tokens. But users have some overheads in managing dedicated passwords or tokens. Accordingly, the biometric identification is delicately introduced to start the era of password-free. However, some disadvantages make these access control schemes still incapable of being effective in real-time online services.

1) They are not *nonintrusive*. Users have to spend extra time in the authentication.

2) They are not *continuous*. The defending system will fail to take further protection once the access control is broken. Behavior-based suspicious account detection is a highly anticipated solution to pursue a nonintrusive and continuous identity authentication for online services. It depends on capturing users' suspicious behavior patterns to discriminate the suspicious accounts. The problem can be divided into two categories: fake/sybil account detection and compromised account detection . The fake/Sybil account's behaviors usually do not conform to the behavioral pattern of the majority. Meantime, the compromised account usually behaves in a pattern that does not conform to its previous one, even behaves like fake/sybil accounts. It can be solved by capturing *mutations* of users' behavioral patterns.Comparing with detecting compromised accounts, detecting fake/sybil accounts is relatively easy since the latter's behaviors are generally more detectable than the former's. It has been extensively studied and can be realized by

various population-level approaches, e.g., clustering classification and statistical or empirical rules . Thus, we *only* focus on the compromised account detection, commonly called *identity theft detection*, based on individual-level behavioral models. Recently, researchers have proposed the individual-level identity theft detection methods by using suspicious behavior detection. The efficacy of these methods significantly depends on the sufficiency of behavior records. They are usually suffering from the low-quality of behavior records due to data collecting limitations or some privacy issues . In particular, when a method only utilizes a specific dimension of behavioral data, the efficacy damaged by poor data is possibly enlarged and the scope of application is limited. Unfortunately, many existing works just concentrate on a specific dimension of users' behavior, such as keystroke click stream touch-interaction and user generated content (UGC) . In this article, we propose an approach to detect identity theft by using multidimensional behavioral records which are possibly insufficient in each dimension. According to such characteristics, we choose the online social network (OSN) as a typical scenario where most users' behaviors are coarsely recorded. In the Internet era, users' behaviors are composited by offline behaviors, online behaviors, social behaviors, and perceptual/cognitive behaviors. The behavioral data can be collected in many applications, such as offline check-ins in location-based services (LBSs), online tips-posting in instant messaging services, and social relationship-making in online social services. Accordingly, we design our method based on users' composite behaviors by these categories In OSNs, user behavioral data that can be used for online identity theft detection are often too low-quality or restricted to build qualified behavioral models due to the difficulty of data collection, the requirement of user privacy, and the fact that some users have a few several behavioral records. We devote ourselves to proving that a high-quality (effective, quick response, and robust) behavioral model can be obtained by integrally using multidimensional behavioral data, even though the data is extremely insufficient in each dimension. Generally, there are two paradigms to integrate behavioral data: the *fused* and *joint* manners.

Fused models are a relatively simple and straightforward kind of composite behavior models (CBMs). They first capture features in each behavior space and then make a comprehensive metric based on these features in different dimensions. With the possible complementary effect among different behavior spaces, they can act as a feasible solution for integration [7], [17]. However, the identification efficacy can be further improved, since fused models neglect potential links among different spaces of behaviors. We take an example where a person posted a picture in an OSN when he/she visited a park. If this composite behavior is simply separated into two independent parts: he/she once posted a picture and he/she once visited a park, the difficulty in relocating him/her from a group of users is possibly increased, since there are more users satisfy these two simple conditions comparing to the original condition. In contrast, the joint model can sufficiently exploit the correlations between behaviors in different dimensions, then increases the certainty of users' behavior patterns, which contributes to a better identification efficacy. The underlying logic for the difference between the joint and fused models can be also explained by the well-known *Chain Rule for Entropy* , which indicates that the entropy of multiple simultaneous events is no more than the sum of the entropies of each individual event, and are equal if the events are independent. It shows that the joint behavior has lower uncertainty comparing to the sum of the uncertainty in each component

Therefore, to fully utilize potential information in composite behaviors for user profiling, we propose a *joint* model, specifically, a joint probabilistic generative model based on Bayesian networks called CBM. It offers a composition of the typical features in two different behavior spaces: check-in location in offline behavior space and UGC in online behavior space. Considering the composite behavior of a user, we assume that the generative mechanism is as follows. When a user plans to visit a venue and simultaneously post tips online, he/she subconsciously selects a specific behavioral pattern according to his/her behavioral distribution. Then, he/she comes up with a topic and a targeted venue based on the present pattern's topic and venue distributions, respectively. Finally, his/her comments are generated following the corresponding topic-word distribution. To estimate the parameters of the mentioned distributions, we adopt the collapsed Gibbs sampling . Based on the joint model CBM, for each composite behavior, denoted by a triple-tuple *(u, v,D)*, we can calculate the chance of user *u* visiting venue *v* and posting a tip online with a set of words *D*. Taking into account different levels of activity of different users, we devise a *relative anomalous score Sr* to measure the occurrence rate of each composite behavior *(u, v,D)*. By these approaches, we finally realize real-time detection (i.e., judging by only one composite behavior) for identity theft suspects We evaluate our joint model by comparing it with three typical models and their fused model [17] on two real-world OSN datasets: Foursquare [43] and Yelp [44]. We adopt the area under the receiver operating characteristic curve (AUC) as the detection efficacy. Particularly, the *recall* [true positive rate (TPR)] reaches up to 65.3% in Foursquare and 72.2% in Yelp, respectively, with the corresponding *disturbance rate* [false-positive rate (FPR)] below 1%, while the fused model can only achieve 60.8% and 60.4% in the same condition, respectively. Note that this performance can be achieved by examining only one composite behavior per authentication, which guarantees the low response latency of our detection method. As an insightful result, we learn that the complementary effect does exist among different dimensions of low-quality records for modeling users' behaviors.

The main contributions are summarized into three folds.

1) We propose a joint model, CBM, to capture both online and offline features of a user's composite behavior to fully exploit coarse behavioral data.

2) We devise a relative anomalous score *Sr* to measure the occurrence rate of each composite behavior for realizing real-time identity theft detection.

3) We perform experiments on two real-world datasets to demonstrate the effectiveness of CBM. The results show that our model outperforms the existing models and has the low response latency.

2.LITERATURE SURVEY

Recently, researchers found that users' behavior can identify their identities [3],

[28], [52]. Typically, behavior-based user identification include two phases: user profiling and user identifying: User profiling is a process to characterize a user with his/her history behavioral data. Some works focus on statistical characteristics to establish the user profile. Naini et al. [53] 10 studied the task of identifying the users by matching the histograms of their data in the anonymous dataset with the histograms from the original dataset. Egele et al. [8] proposed a behavior-based method to identify compromises of highprofile accounts. Ruan et al. [30] conducted a study on online user behavior by collecting and analyzing user clickstreams of a well known OSN. Lesaege et al. [29] developed a topic model extending the Latent Dirichlet Allocation (LDA) to identify the active users. Viswanath et al. [44] presented a technique based on Principal Component Analysis (PCA) that accurately modeled the "like" behavior of normal users in Facebook and identified significant deviations from it as anomalous behaviors. Tsikerdekis and Zeadally [54] presented a detection method based on nonverbal behavior for identity deception, which can be applied to many types of social media. These methods above mainly concentrated on a specific dimension of the composite behavior without utilizing the correlations among multi-dimensional behavior data. Vedran et al. [55] explored the complex interaction between social and geospatial behavior and demonstrated that social behavior can be predicted with high precision. Yin et al. [36] proposed a probabilistic generative model combining use spatiotemporal data and semantic information to predict user's behavior. These studies implied that composite behavior features are possibly helpful for user identification. User identifying is a process to match the same user in two datasets or distinguish anomalous users/behaviors. User identifying can be applied to a variety of tasks, such as detecting anomalous users or match users across different data sources. Mazzawi et al. [56] presented a novel approach for detecting malicious user activity in databases by checking user's self-consistency and global-consistency. Lee and Kim [32] proposed a suspicious URL detection system for Twitter to detect users' anomalous behaviors. Cao et al. [22] designed and implemented a malicious account detection system for detecting both

fake and compromised real user accounts. Zhou et al. [57] proposed an FRUI algorithm to match users among multiple OSNs. These works mainly detected the populationlevel anomalous behaviors which indicated strongly difference from other behaviors. While, they did not consider that the individual-level coherence of users' behavioral patterns can be utilized to detect online identity thieves.

3. EXISTING SYSTEM

Sitova*et al.* introduced hand movement, orientation, and grasp (HMOG), a set of behavioral features to continuously authenticate smartphone users. Rajoub and Zwiggelaar used thermal imaging to monitor the periorbital region's thermal variations and test whether it can offer a discriminative signature for detecting deception. However, these biometric technologies usually require expensive hardware devices which makes it inconvenient and difficult to popularize.

Abouelenien*et al.* explored a multimodal deception detection approach that relied on a novel dataset of 149 multimodal recordings, and integrated multiple physiological, linguistic, and thermal features. These works indicated that users' behavior patterns can represent their identities. Many studies turn to utilize users' behavior patterns for identifications. Behavior-based methods were born at the right moment, which plays important roles in a wide range of tasks including preventing and detecting identity theft. Typically, behavior-based user identification includes two phases: user profiling and user identifying.

User profiling is a process to characterize a user with his/her history behavioral data. Some works focus on statistical characteristics, such as the mean, variance, median, or frequency of a variable, to establish the user profile. Naini *et al.* studied the task of identifying the users by matching the histograms of their data in the anonymous dataset with the histograms from the original dataset. But it mainly relied on experts' experience since different cases usually have different characteristics.

Egele*et al.*proposed a behavior-based method to identify compromises of individual high-profile accounts. However, it required high-profile accounts which were difficult to obtain. Other researchers discovered other features, such as tracing patterns, topic and spatial distributions, to describe user identity. Ruan *et al.* conducted a study on online user behavior by collecting and analyzing user clickstreams of a well-known OSN. Lesaege*et al.* developed a topic model extending the LDA to identify the active users. Viswanath*et al.*presented a technique based on principal component analysis (PCA) that accurately modeled the "like" behavior of normal users in Facebook and identified significant deviations from it as anomalous behaviors. Zaeem *et al.* proposed an approach that involved the novel collection of online news stories and reports on the topic of identity theft. Lichman and Smyth [48] proposed MKDE model to accurately characterize and predict the spatial pattern of an individual's events.

Tsikerdekis and Zeadally presented a detection method based on nonverbal behavior for identity deception, which can be applied to many types of social media. These methods above mainly concentrated on a specific dimension of the composite behavior and seldom thought about utilizing multidimensional behavior data. Sekara*etal.*explored the complex interaction between social and geospatial behavior and demonstrated that social behavior can be predicted with high precision. It indicated that composite behavior features can identify one's identity. Yin *etal.*proposed a probabilistic generative model combining the use of spatiotemporal data and semantic information to predict user's behavior. Nilizadeh*etal.*presented POISED, a system that leverages the differences in propagation between benign and malicious messages on social networks to identify spam and other unwanted content. These studies implied that composite behavior features are possibly helpful for user identification.

Disadvantages

1) LDA model performs poorly in both datasets which may indicate its performance is strongly sensitive to the data quality.

2) CF-KDE and LDA model performs not well in Yelp dataset comparing to Foursquare dataset, but the fused model [17] observes a surprising reversion.

3) The joint model based on *relative anomalous score Sr* outperforms the model based on *logarithmic anomalous score Sl*.

4) The joint model (i.e., JOINT-SR, the joint model in the following content of the system

all refer to the joint model based on *Sr*) is indeed superior to the fused model.

3.1 PROPOSED SYSTEM

In this article, we propose an approach to detect identity theft by using multidimensional behavioral records which are possibly insufficient in each dimension. According to such characteristics, we choose the online social network (OSN) as a typical scenario where most users' behaviors are coarsely recorded [39]. In the Internet era, users' behaviors are composited by offline behaviors, online behaviors, social behaviors, and perceptual/cognitive behaviors. The behavioral data can be collected in many applications, such as offline check-ins in location-based services (LBSs), online tips-posting in instant messaging services, and social relationship-making in online social services. Accordingly, we design our method based on users' composite behaviors by these categories.

In OSNs, user behavioral data that can be used for online identity theft detection are often too low-quality or restricted to build qualified behavioral models due to the difficulty of data collection, the requirement of user privacy, and the fact that some users have a few several behavioral records. We devote ourselves to proving that a high-quality (effective, quickresponse, and robust) behavioral model can be obtained by integrally using multidimensional behavioral data, even though the data is extremely insufficient in each dimension.

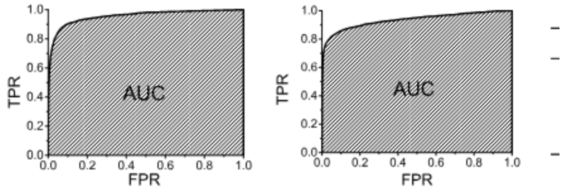
Advantages

1) We propose a joint model, CBM, to capture both online and offline features of a user's composite behavior to fully exploit coarse behavioral data.

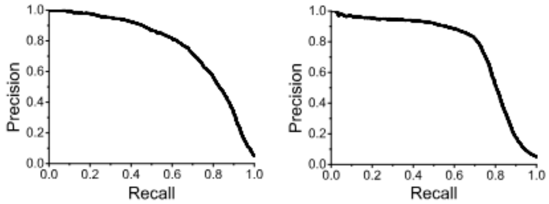
2) We devise a relative anomalous score Sr to measure the occurrence rate of each composite behavior for realizing real-time identity theft detection.

3) We perform experiments on two real-world datasets to demonstrate the effectiveness of CBM. The results show that our model outperforms the existing models and has the low response latency.

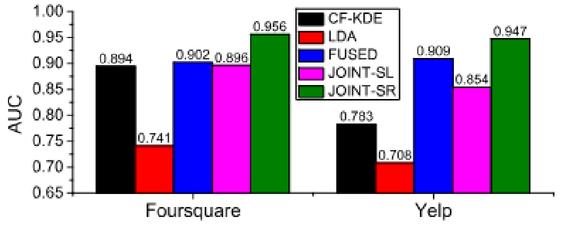
4. OUTPUTSCREENS

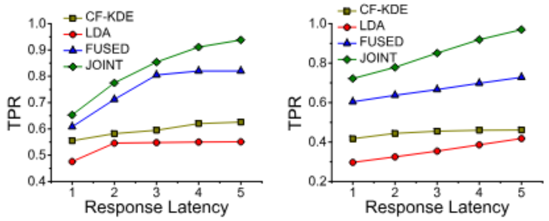


(a) Foursquare (b) Yelp

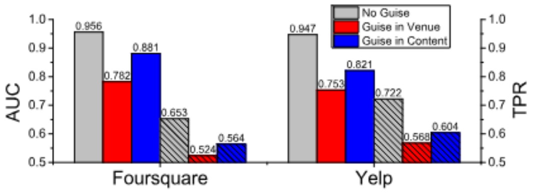


(a) Foursquare (b) Yelp





(a) Foursquare (b) Yelp



5. CONCLUSION

In this paper, we pose the problem of fraudulent insurance claim identification as a feature generation and classification process. We formulate the problem over a minimal, definitive claim data consisting of procedure and diagnosis codes, because accessing richer datasets are often prohibited by law and present inconsistencies among different software systems. We introduce clinical concepts over procedure and diagnosis codes as a new representation learning approach. We assume that every claim is a representation of latent or obvious Mixtures of Clinical Concepts which in turn are mixtures of diagnosis and procedure codes. We extend the MCC model using Long-Short Term Memory network (MCC + LSTM) and Robust Principal Component Analysis (MCC + RPCA) to filter the significant

concepts from claims and classify them as fraudulent or non fraudulent. Our results demonstrate an improvement scope to find fraudulent healthcare claims with minimal information. Both MCC and MCC + RPCA exhibit consistent behavior for varying concept sizes and replacement probabilities in thenegative claim generation process. MCC + LSTM reachesan accuracy, precision, and recall scores of 59%, 61%, and50%, respectively on the inpatient dataset. Besides, it presents78%, 83%, and 72% accuracy, precision, and recall scores, respectively on the outpatient dataset. We notice similarity between the results of MCC and MCC + RPCA, as both use an SVM classifier. We believe that the proposed problem formulation, representation learning and solution will initiate new research on fraudulent insurance claim detection using minimal, but definitive data.

6. REFERENCE

1. Government of Pakistan. Introduction, Sehat Sahulat Program.
2019. https://sehatinsafcard.com/introduction.php. Accessed January 2023.
2. Government of Pakistan. Benefits Package. 2019.
https://sehatinsafcard.com/benefits.php. Accessed January 2023.
3. Government of United States. Centers for Medicare and Medicaid Services. 1965.
https://www.medicare.gov/. Accessed January 2023.
4. Gee J, Button M, Brooks G. The financial cost of healthcare fraud: what data from around the world shows. 2010.
5. Berwick DM, Hackbarth AD. Eliminating waste in US health care.
JAMA. 2012;**307**(14):1513–1516.
doi: 10.1001/jama.2012.362. [PubMed] [CrossRef] [Google Scholar]
6. M King K. Progress Made, but More Action Needed to Address Medicare Fraud, Waste, and Abuse. 2014.
https://www.gao.gov/assets/gao-14-560t.pdf. Accessed January 2023.