Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

COMPOSITE BEHAVIOURAL MODELLING AS AN APPROACH TO THE IDENTIFICATION OF IDENTITY THEFT IN ONLINE SOCIAL NETWORKS

[1]MANOJ KUMAR ILLURI,[2]H.MADHUSUDHANA RAO
[1]Student,[2]Assistant professor, MCA,M.Phill,(Ph.D)
Department of CSE

ABSTARCT:

In the work that we are presenting here, our objective is to construct a bridge from coarse behavioural data to an efficient, rapid-response, and reliable behavioural model for the detection of online identity theft. We focus on this problem in the context of online social networks (OSNs), where individuals often have composite behavioural records that are made up of multidimensional low-quality data, such as offline check-ins and online user-generated content (UGC). We validate that there is a complementing impact among various dimensions of data for modelling the behavioural patterns of users, which is an interesting discovery that we came up with.

We suggest a joint model (instead of a fused model) to capture both online and offline aspects of a user's composite behaviour in order to fully leverage such a complementing impact. This will allow us to truly capitalise on the effect. We assess the proposed joint model by contrasting it with usual models and their fused model on two real-world datasets, namely Foursquare and Yelp. These datasets include information gathered from the actual world. The experimental findings demonstrate that our model works better than the ones that are already in use, with values of 0.956 and 0.947 for the area under the receiver operating characteristic curve (AUC) when applied to Foursquare and Yelp, respectively. In particular, the recall (true positive rate) may reach up to 65.3% in Foursquare and 72.2% in Yelp, while the equivalent disturbance rate (false-positive rate) remains below 1% on each of these platforms.

It is important to note that these results may be obtained by analysing a single composite behaviour, which ensures that our approach has a short reaction latency. This fact is worthy of remark. This project would provide the cybersecurity community with fresh insights on whether

or whether real-time online identity authentication may be enhanced via modelling individuals' composite behavioural patterns, as well as how such an improvement can be implemented.

Composite behavioural modelling, identity theft detection, joint modelling, and online social networks (OSNs) are some of the terms included in this index.

I.INTRODUCTION:

WITH the rapid development of the Internet, more and more affairs, e.g., mailing [1], health caring [2], shopping [3], booking hotels, and purchasing tickets, are handled online [4]. Meanwhile, the Internet also brings sundry potential risks of invasions, such as losing financial information [5], identity theft [6], and privacy leakage [3]. 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 [7], 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 Manuscript received March 11, 2020; revised May 10, 2021; accepted June 9, 2021. The work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61972287, in part by the Major Project of the Ministry of Industry and Information Technology of China under Grant TC200H01J, and in part by the Municipal Human Resources Development Program for Outstanding Young Talents in Shanghai. (Corresponding author: Cheng Wang.) The authors are with the Key Laboratory of the Ministry of Education for Embedded System and Service Computing, Department of Computer Science, Tongji University, Shanghai 201804, China (e-mail: chengwang@tongji.edu.cn; 1830823@tongji.edu.cn; boyang@tongji.edu.cn). Digital Object Identifier 10.1109/TCSS.2021.3092007 accounts are usually the portals of most cybercrimes [1], such as blackmail [5], fraud [8], and spam [9], [10]. Thus, identity theft detection is essential to guarantee users' security in the cyberworld. Traditional identity authentication methods are mostly based on access control schemes, e.g., passwords and tokens [11], [12]. But users have some overheads in managing dedicated passwords or tokens. Accordingly, the biometric identification [13]–[15] 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 [16], [17]. 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 [16], [18], [19] 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 [20] and compromised account detection [21]. 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 [22], [23], classification [5], [24]–[26] and statistical or empirical rules [8], [27], [28]. 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 [9], [29]–[35]. 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 [3]. 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 [29], clickstream [32], [36], touch-interaction [37], and user generated content (UGC) [9], [33], [34], [38]. 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. 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 [40], 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 [41]. 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 [42]. 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. The rest of this article is organized as follows. We give an overview of our solution in Section II. Then, we present our method in Section III, and make the validation in Section IV. We provide a literature review in Section V. Finally, we draw conclusions in Section VI.

II.LITERATURE SURVEY

Online identity theft occurs when a thief steals a user's personal data and impersonates the user's account. Generally, a thief usually first gathers information about a targeted user to steal his/her identity and then use the stolen identity to interact with other people to get further benefits [4]. Criminals in different online services usually have different motivations. An OSN user's behavior is usually composed of online and offline behaviors occurring in different behavioral spaces [17]. Based on this fact, we aim at devising a joint model to embrace them into a unified model to deeply extract information. Before presenting our joint model, named CBM, we provide some conceptions as the preparations. The relevant notations are listed in Table I. Definition 1 (Composite Behavior): A composite behavior, denoted by a four-tuple (u,v, D,t), indicates that at time t, user u visits venue v and simultaneously posts a tip consisting of a set of words D online. In this work, the representation of a composite behavior can be simplified into a triple-tuple (u,v, D). We remark that for a composite behavior, the occurring time t is a significant factor. Two types of time attributes play important roles in digging potential information for improving the identification. The first is the sequential correlation of behaviors. However, in some OSNs, the time intervals between adjacent behavioral records are usually overlong, which leads that the sequential correlations cannot be captured effectively. The second is the temporal property of behaviors, e.g., periodicity and preference variance over time. However, in some OSNs, the occurring time is recorded with a low resolution, e.g., by day, which shields the possible dependency of a user's behavior on the occurring time. Thus, it is difficult to obtain reliable time-related features of users' behaviors. Since we aim to propose a practical method based on uncustomized datasets of user behaviors, we only concentrate on the dependency between a user's check-in location and tip-posting content of each behavior, taking no account of the impact of specific occurring time in this work. Thus, the representation of a composite behavior can be simplified into a triple-tuple (u,v, D) without confusion in this article. The graphical representation of our joint model CBM is demonstrated in Fig. 1. Our model is mainly based on the following two assumptions. 1) Each user behaves in multiple patterns with different possibilities. 2) Users with similar behavioral patterns have similar interests in topics and places. To describe the features of users' behaviors, we first introduce the topic of tips. Definition 2 (Topic, [45]): Given a set of words W, a topic z is represented by a multinomial distribution over words, denoted by φ_z, whose each component $\varphi_{z,w}$ denotes the probability of

word w occurring in topic z. Next, we formulate a specific behavioral pattern of users by a conception called community. Definition 3 (Community): A community is a set of users with the same behavioral pattern. Let C denote the set of all communities. A community c ∈ C has two critical parameters. 1) A topic distribution θc, whose component, say θc,z, indicates the probability that the users in community c send a message with topic z. 2) A spatial distribution ϑc, whose component, say ϑc,v , represents the chance that users in community c visit venue v. More specifically, we assume that a community is formed by the following procedure. Each user u is included in communities according to a multinomial distribution, denoted by πu. That is, each component of πu, say πu,c, denotes u's affiliation degree to community c. Similarly, we allocate each community c with a topic distribution θc to represent its online topic preference and a spatial distribution ϑc to represent its offline mobility pattern.

Generally, users take actions according to their regular behavioral patterns which are represented by the corresponding communities (Definition 3). We present the behavioral generative process in Algorithm 1: When a user u is going to visit a venue and post online tips there, he/she subconsciously selects a specific behavioral pattern, denoted by community c, according to his/her community distribution πu (line 11). Then, he/she comes up with a topic z and a targeted venue v based on the present community's topic and venue distributions (θc and ϑc, respectively) (lines 12 and 13). Finally, the words of his/her tips in D are generated following the topic-word distribution φz (line 15). Exact inference of our joint model CBM is difficult due to the intractable normalizing constant of the posterior distribution [42].

III.CONCLUSION AND FUTURE WORK

We investigate the feasibility of building a ladder from low-quality behavioral data to a high-performance behavioral model for user identification in OSNs. By deeply exploiting the complementary effect among OSN users' multidimensional behaviors, we propose a joint probabilistic generative model by integrating online and offline behaviors. When the designed joint model is applied to identity theft detection in OSNs, its comprehensive performance, in terms of the detection efficacy, response latency, and robustness, is validated by extensive evaluations on real-life OSN datasets. Particularly, the joint model significantly outperforms the existing fused model. Our behavior-based method mainly aims at detecting identity thieves after

the access control of the account is broken. Then, it is easy and promising to incorporate our method into traditional methods to solve the identity theft problem better.

ACKNOWLEDGMENT

The authors sincerely thank the anonymous referees for their helpful comments and suggestions.

REFERENCES

[1] J. Onaolapo, E. Mariconti, and G. Stringhini, "What happens after you are pwnd: Understanding the use of leaked Webmail credentials in the wild," in Proc. Internet Meas. Conf., Nov. 2016, pp. 65–79.

[2] A. Mohan, "A medical domain collaborative anomaly detection framework for identifying medical identity theft," in Proc. Int. Conf. Collaboration Technol. Syst. (CTS), May 2014, pp. 428–435.

[3] Y.-A. de Montjoye, L. Radaelli, V. K. Singh, and A. S. Pentland, "Unique in the shopping mall: On the reidentifiability of credit card metadata," Science, vol. 347, no. 6221, pp. 536–539, Jan. 2015.

[4] P. Hyman, "Cybercrime: It's serious, but exactly how serious?" Commun. ACM, vol. 56, no. 3, pp. 18–20, Mar. 2013.

[5] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda, "All your contacts are belong to us: Automated identity theft attacks on social networks," in Proc. 18th Int. Conf. World Wide Web (WWW), 2009, pp. 551–560. [6] J. Lynch, "Identity theft in cyberspace: Crime control methods and their effectiveness in combating phishing attacks," Berkeley Technol. Law J., vol. 20, no. 1, pp. 259–300, 2005.

[7] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna, "Towards detecting compromised accounts on social networks," IEEE Trans. Dependable Secure Comput., vol. 14, no. 4, pp. 447–460, Jul. 2017.

[8] T. C. Pratt, K. Holtfreter, and M. D. Reisig, "Routine online activity and Internet fraud targeting: Extending the generality of routine activity theory," J. Res. Crime Delinquency, vol. 47, no. 3, pp. 267–296, Aug. 2010.

[9] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song, "Design and evaluation of a real-time URL spam filtering service," in Proc. IEEE Symp. Secur. Privacy, May 2011, pp. 447–462.

[10] H. Li et al., "Bimodal distribution and co-bursting in review spam detection," in Proc. 26th Int. Conf. World Wide Web, Apr. 2017, pp. 1063–1072.