On Complementary Effect of Blended Behavioral Analysis for Identity Theft Detection in Mobile Social Networks

Cheng Wang(✉), Jing Luo, Bo Yang, and Changjun Jiang

Key Laboratory of Embedded System and Service Computing, Ministry of Education, Department of Computer Science and Technology, Tongji University, Shanghai, China
{chengwang,jingluo,boyang,cjjiang}@tongji.edu.cn

Abstract. User behavioral analysis is expected to act as a promising technique for identity theft detection in the Internet. The performance of this paradigm extremely depends on a good individual-level user behavioral model. Such a good model for a specific behavior is often hard to obtain due to the insufficiency of data for this behavior. The insufficiency of specific data is mainly led by the prevalent sparsity of users' collectable behavioral footprints. This work aims to address whether it is feasible to effectively detect identify thefts by jointly using multiple unreliable behavioral models from sparse individual-level records. We focus on this issue in mobile social networks (MSNs) with multiple dimensions of collectable but sparse data of user behavior, i.e., making check-ins, posing tips and forming friendships. Based on these sparse data, we build user spatial distribution model, user post interest model and user social preference model, respectively. Here, as the arguments, we validate that there is indeed a complementary effect in multi-dimensional blended behavioral analysis for identity theft detection in MSNs.

Keywords: Mobile social networks · Identity theft detection
Blended behavioral analysis · Complementary effect

1 Introduction

Mobile Internet has been increasingly gaining popularity. To say that a global phenomenon is almost an understatement, as the number of worldwide social network users is expected to grow to around 2.5 billion in 2018, around a third of Earth's entire population. Projections also show that by 2018, over 75% of Facebook users worldwide will access the service through their mobile phone[1].

However, due to the ever increasing number of users and the advent of information boom, people's property is also rapidly digitized (private photos, customer sensitive data, intellectual property, etc.). According to a new survey

[1] One of the largest statistics portals, http://www.statista.com/.

L. Zhu and S. Zhong (Eds.): MSN 2017, CCIS 747, pp. 32–44, 2018.
https://doi.org/10.1007/978-981-10-8890-2_3

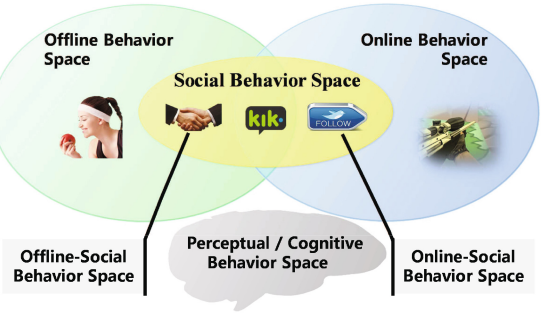

Fig. 1. Modern human behaviors usually take place in multiple spaces, such as the physical space of reality, virtual cyber space and social space.

from security software company Webroot[2], MSNs users are more likely to face security threats such as the loss of financial information, identity theft, and the infringement of the right to privacy. A more common strategy is to hijack a user account, then the hackers send messages to their contacts and deceive them to send money to a "friend in trouble". Because social network is mostly acquaintance network, users can easily relax the vigilance and eventually be deceived. When a user logs in to a social network, he or she will generate a sequence of behaviors. By modeling these behavioral habits, it is possible to derive behavioral characteristics that uniquely discern the user's identity. Studies have shown that private traits and attributes are predictable from digital records of human behavior and computer-based personality judgments are more accurate than those made by humans. Montjoye et al. [1] investigated three months of credit card records for 1.1 million people and showed that four spatiotemporal points are enough to uniquely reidentify 90% of individuals. It strongly proved that the behavior of the track can uniquely identify a person.

The research of user behavior has caused widespread concern in recent years. In a perspective, human beings live in a multiple space, such as the realistic physical space, virtual cyber space and social space. Mobile Internet is essentially to provide users with real-time switching channels. The behavior can be divided into offline behavior in the physical space, online behavior in the cyber space and social behavior in the social space. These form a blended space consisting of offline behavior space, online behavior space and social behavior space, which is illustrated in Fig. 1.

In this work, we grasp the essence of Mobile Internet serving as a portal for users among multiple spaces, e.g. physical and cyber spaces, which makes user behavior blended by multiple dimensional. The main issue to be addressed is if it is feasible to overcome the sparsity of behavioral data by cooperatively using multiple dimensional behavior. More specifically, we focus on behavioral data in three dimensions, i.e., check-ins, tips and friendships. The identity theft detection can be carried out by the comprehensive analysis of the above three behavior

[2] The largest privately held cybersecurity organization based in the USA, operating globally across North America, EMEA and APAC, https://www.webroot.com/.

Table 1. Notations

Symbol	Meaning
I_ξ	The set of users who can be identified in each dimension, $\xi \in$ { DoC, DoT, DoF, DoCT, DoCF, DoTF, DoCTF }
S_{log}^{che}	The threshold used to determine user identity in the method of DoC
D_{JS}^{tip}	The threshold used to determine user identity in the method of DoT
D_{JS}^{fri}	The threshold used to determine user identity in the method of DoF
n_{vow}^{λ}	The minimum number of valid words contained in corresponding method, $\lambda \in$ { tip, fri }
θ_{his}^{λ}	Vector of topic probability distribution, which indicates the behavior characteristic of user historical data, $\lambda \in$ { tip, fri }
θ_{new}^{λ}	Vector of topic probability distribution, which indicates the behavior characteristic of user newly generated data, $\lambda \in$ { tip, fri }

data. We perform the detection on two Location-based Social Networks (LBSNs) data sets [2], i.e., Foursquare and Yelp. We choose several metrics to evaluate the detection performance. One is the *intercept rate* (IR), that is, the proportion of anomalous accounts that are intercepted accurately. The other is the *disturb rate* (DR), which is the proportion of the accounts that are erroneously intercepted to be anomalous in the normal account. In summary, this paper makes the following contributions:

- We propose three detection methods via user behavior for identity theft detection.
- We obtain the performance of user behavior models for identity theft detection on two real-life data sets.
- We validate that there is indeed a complementary effect on multi-dimensional blended behavioral analysis for identity theft detection in MSNs.

2 Problem Description and Settings

Identity theft refers to somebody stealing your personal data and impersonating you to, for example, shop under your name. Statistics showed that six percent of users stated that they suffered from identity theft. In this work, we propose the approach for identity theft detection via user blended behavior in MSNs. We mainly address two issues: For one thing we aim at building the behavior model for each dimension and proposing the corresponding detection method to effectively identify the user authenticity; for another the poor detection performance causing by data sparsity should also be solved due to the complementary effect of the blended behavioral analysis.

Our first object is to build the user's behavior model based on their generated data. User behavior data are the check-in records at the location, online text content, and user-generated social relationship. We build three user behavior models corresponding to each dimension. Corresponding to our work, we call

them three dimensions, namely dimension of check-ins (DoC), dimension of tips (DoT) and dimension of friendships (DoF). In Table 1, we give an explanation of the symbols appearing in this paper. Due to various constraints, the behavior data are sparse for a particular user in single dimension. Since the detection is done by the individual-level user behavioral model, we define those who have sufficient data and can be identified by our method of this dimension as *Identifiable Users*. In each model, we need to find the appropriate criteria to distinguish between the user's own and non-users themselves, i.e., S_{log}^{che}, D_{JS}^{tip} and D_{JS}^{fri}. In order to achieve identity theft detection, we chose three metrics to evaluate the detection performance, which are *intercept rate* (IR), *disturb rate* (DR) and *precision rate* (PR), respectively. They are defined as follows:

$$\text{Intercept Rate} = \frac{\#\text{ users intercepted correctly}}{\#\text{ all anomalous users}}, \tag{1}$$

$$\text{Disturb Rate} = \frac{\#\text{ users intercepted wrongly}}{\#\text{ all normal users}}, \tag{2}$$

$$\text{Precision Rate} = \frac{\#\text{ users intercepted correctly}}{\#\text{ all intercepted users}}. \tag{3}$$

3 Detection Method

In this section, we present the user behavior models for each dimension and the methods for identity theft detection via the corresponding models.

3.1 Detection via Check-Ins

Modeling User Spatial Distributions. We model the user spatial distribution (USDM) by using the method based on the kernel density estimation. According to the history data of the users, the probability density function of each user is obtained by using the mixed kernel density estimation (MKDE) method [3]. Let $E = \{e^1, ..., e^n\}$ be a set of historical events where $e^j =< x, y >$ is a two-dimensional spatial location, $1 \leq j \leq n$, and where we have suppressed any dependence on individual i for the moment and dropped dependence on time t. Let E denote the training data set. A simple method for estimating a bivariate density function from such kind of data is to use a single fixed 2×2 bandwidth matrix H and a Gaussian kernel function $K(\cdot)$. More specifically, we assume the bandwidth matrix $H = \begin{pmatrix} h & 0 \\ 0 & h \end{pmatrix}$. This results in a bivariate KDE of the following form:

$$f_{KD}\left(e|E,h\right) = \frac{1}{n}\sum_{j=1}^{n} K_h\left(e - e^j\right), \tag{4}$$

$$K_h\left(x\right) = \frac{1}{2\pi h}\exp\left(-\frac{1}{2}x^T H^{-1} x\right), H = \begin{pmatrix} h & 0 \\ 0 & h \end{pmatrix}, \tag{5}$$

where e is the location for which we would like to calculate the probability density, and $h > 0$ is a fixed scalar bandwidth parameter for all events in E.

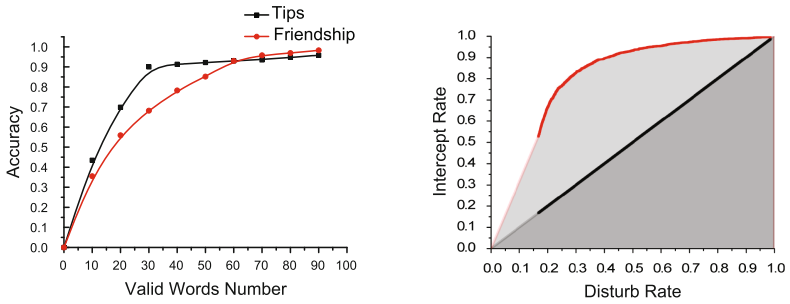

Fig. 2. (a) is the number of valid words for identification accuracy about the DoT and DoF, respectively and (b) is the ROC curve of identity theft detection through the social topology. (Color figure online)

To deal with "cold start" problem, we use a mixture model as follows:

$$f_{MKD}\left(e|E,h\right) = \alpha f_{KD}\left(e|E_1\right) + \left(1-\alpha\right) f_{KD}\left(e|E_2\right), \tag{6}$$

where E_1 is a set of an individual's historical events (individual component), E_2 is a set of the friends' historical events (friendship component) and α is the weight for individual component.

Identity Theft Detection. We compute the average log-likelihood of each user i's check-in records in the test set:

$$S_i = -\frac{1}{n_i}\sum_{r=1}^{n_i} \log f_{MKD_i}\left(e^r|E,h_i\right), \tag{7}$$

where n_i is the number of user i's check-in records in the test set. The larger S_i, the more likely the records are conducted by himself. The key is to set the anomalous threshold, and define it as S_{log}^{che}. When the user i's S_i is greater than the baseline, we consider the account to be anomalous.

3.2 Detection via Tips

Modeling User Post Interests. We build the user post interest model (UPIM) by using the user-generated tips in the online behavior space. Each user's historical tips accumulate as a document, and then all the user's tips constitute a large-scale corpus. In LDA [4], each document may be viewed as a mixture of various topics. Through the document generation model, we can get the topic probability distribution of the document. In fact, the topic probability distribution corresponding to each document is the interest probability distribution of the user. We define the vector of topic probability distribution generated by the historical data as θ_{his}^{tip}.

Identity Theft Detection. After the new generated tips are processed, the vector of topic probability distribution can be calculated as θ_{new}^{tip}. We count the number of words assigned to kth topic, and denote it as $n(k)$. α is a hyperparameter, and we set $\alpha = 0.5$. The kth component of the topic proportion vector can be computed as:

$$\theta_{new} = \frac{n\left(k\right) + \alpha}{\sum_{i=1}^{K}\left(n\left(i\right) + \alpha\right)}. \tag{8}$$

In order to ensure that the tips can get a reasonable topic probability distribution, we hope that the accumulation of tips as much as possible. However, considering the efficiency of the model, we need to be able to give the identity judgment in a short time. We solve this problem by experimentally training the number of valid words. Then we define the minimum number of valid words contained in the tips as n_{vow}^{tip}. Through experiments shown in Fig. 2(a), we find that when the number of valid words reaches 30, the accuracy tends to be stable and very impressive, so we set the parameter n_{vow}^{tip} to 30 for the DoT. The interest probability distribution of tips for the user's new post is calculated as θ_{new}^{tip} by using Eq. 8.

For the discrete probability distributions P and Q, the Kullback-Leibler divergence from Q to P is defined as:

$$D_{KL}\left(p, q\right) = \sum_{i=1}^{T} p_i \cdot \ln \frac{p_i}{q_i}. \tag{9}$$

An alternative is given via the γ divergence,

$$\begin{aligned} D_{KL}\left(p, q\right) = \gamma D_{KL}\left(p, \gamma \cdot p + \left(1 - \gamma\right) q\right) \\ + \left(1 - \gamma\right) D_{KL}\left(q, \gamma \cdot p + \left(1 - \gamma\right) q\right), \end{aligned} \tag{10}$$

which can be interpreted as the expected information gain about X from discovering which probability distribution X is drawn from, P or Q, if they currently have probabilities γ and $(1 - \gamma)$ respectively.

The value $\gamma = 0.5$ gives the Jensen−Shannon divergence by

$$D_{JS}\left(p, q\right) = \frac{1}{2}\left[D_{KL}\left(p, M\right) + D_{KL}\left(q, M\right)\right], \tag{11}$$

where $M = \frac{p+q}{2}$ is the average of the two distributions.

The similarity of two topic probability distributions can be judged by JS divergence. According to its definition, a smaller divergence of JS indicates a higher similarity of the two distributions. In this work, θ_{his}^{tip} is P, and θ_{new}^{tip} is Q. So what we need to calculate is $D_{JS}^{tip}\left(\theta_{his}^{tip}, \theta_{new}^{tip}\right)$. Set the anomalous threshold to D_{JS}^{tip}. When the $D_{JS}^{tip}\left(\theta_{his}^{tip}, \theta_{new}^{tip}\right)$ of two probability distribution is greater than D_{JS}^{tip}, we consider the account to be an anomalous account.

3.3 Detection via Friendships

Modeling User Social Preferences. We use the social relationships and friend-generated tips to build the user social preferences model (USPM). Generally speaking, when the user social relationship is complete, the friendship topological structure can be utilized to model the friendship community. Due to data imperfection, the topology of friendships cannot be utilized to identify users. Figure 2(b) is the ROC (Receiver Operating Characteristic) curve of identity theft detection via the social topology, which shows an extremely poor performance. Especially when the DR is relatively low, the effect of interception is very poor, which can not be used for identity detection.

It is considered that the user's friendship community tends to be stable. If a user follows the new friends who are very different from previous ones, we would like to believe that the account may be anomalous. Therefore, we need to get the characteristics of friendship community for each user according to user's friendship and user-generated content.

Identity Theft Detection. For the new friends of the user, we cluster the preferences of these new friends for each user. Similar to the DoT, we calculate the $D_{JS}^{fri}\left(\theta_{his}^{fri}, \theta_{new}^{fri}\right)$ by using Eq. 11.

The red curve is the process of training the minimum number of valid words which is defined as n_{vow}^{fri} in Fig. 2(a). Therefore, when n_{vow}^{fri} is 60, it can achieve a stable and satisfactory accuracy for the detection method based on USPM. Set the anomalous threshold to D_{JS}^{fri}, which we define it as the baseline of the method for the DoF. When the values of D_{JS}^{fir} of two probability distributions are greater than the baseline, we consider that the preferences between the new and old friends vary in a wide range and the account could be anomalous.

4 Experiments

In this section, we present the experiments to evaluate our models and validate the complementary effect of blended behavioral analysis for identity theft detection in MSNs.

4.1 Data Sets

We present a large-scale study of user behavior for two real-life datasets of Foursquare and Yelp, and conduct experiments of about 70 thousand users that spans a period of more than 100 days. In the location-based mobile social network, we study the check-in record (DoC), online text information (DoT) and social relationship (DoF) for each user, and model the user behavior from three dimensions. Table 2 shows the number of users and the number of identifiable users in each dimension.

Table 2. Number of identifiable users in each dimension.

Dataset	Users	I_{DoC}	I_{DoT}	I_{DoF}
Foursquare	23537	18624	1062	17472
Yelp	43137	29885	1912	18484

4.2 Theft Simulation

The detection of suspicious accounts can be attributed to two types, respectively fake account detection and compromised account detection. For all accounts, we first consider the population-level suspicious behavior detection. Its purpose is to detect the outlier. If the difference is obvious, the account will be blocked as suspicious account. As for compromised account detection, although it passes the population-level suspicious behavior detection successfully, we do not fully trust it. It needs to be further detected, which is individual-level suspicious behavior detection. That is, even if the user's current behavior is not outlier, we have to detect whether the behavior is the same as himself before. If the difference is obvious, we have to doubt his identity. Our work is to solve the last and most critical issue. It means that all the suspicious behavior we have taken into account.

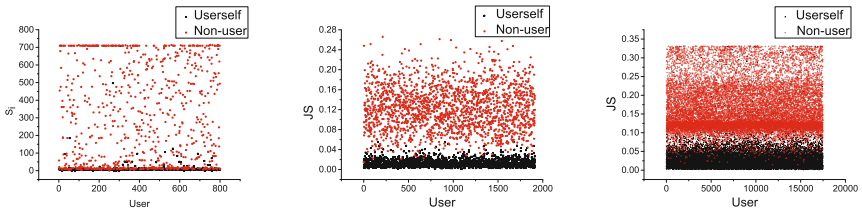

Fig. 3. The experiments of identity theft detection via Check-in, Tips and Friendship. (Color figure online)

It is hard to obtain the real data set marked with identity theft. However, it is not a pain point for our research. For identity theft detection, the main concern is how to simulate identity theft events. Our model presupposes that an attacker who steals an account will immediately start using it as they would have been some other random user of the service. Then the attacker will generate a series of behavioral records with personal characteristics. These behavioral data are those records that we have replaced with others. In fact, it is most difficult to distinguish them in the detection of suspicious behavior. One obvious fact is that if the model can accurately identify these replaced users, that is, simulated thieves, then for those thieves whose behavior is outlier, the detection performance will be more effective.

4.3 Experimental Setup

Firstly, we preprocess the data for each dimension separately according to the above methods. Secondly, we simulate identity theft by randomly replacing part of user's data with others. Thirdly, we determine the threshold of the anomalous account for the methods proposed in Sect. 3. In the DoC, we calculate S_i for each user. In the DoT and DoF, we separately calculate $D_{JS}^{tip}\left(\theta_{his}^{tip}, \theta_{new}^{tip}\right)$ and $D_{JS}^{fri}\left(\theta_{his}^{fri}, \theta_{new}^{fri}\right)$ for each user. Figure 3 is the experimental results on three dimensions. The red scatter is the result of the similarity between the user's new record and the historical data. The black scatter is the result of the similarity between the non-user's new record and the historical data. The results show that these two values are significantly different. Therefore, we can use these differences to carry out identity theft detection.

4.4 Parameter Settings

For identity theft detection, our goal is to improve the IR-DR trade-off, that is, to achieve a high IR via their blended behavior model in the case of a relatively low DR. Figure 4 is the IRs and DRs with the threshold changes in three dimensions, respectively. It can be observed that the DRs will rise sharply when the threshold reaches a certain value. We need to determine the appropriate threshold so that the IRs are as high as possible in the case of tolerable DRs. Therefore, we set the threshold at this critical point for identity anomalies. So, in each dimension, we finally determine the threshold as follows:

(1) In the DoC, we set S_{log}^{Che} to 207.
(2) In the DoT, we set D_{JS}^{tip} to 0.04.
(3) In the DoF, we set D_{JS}^{fri} to 0.08.

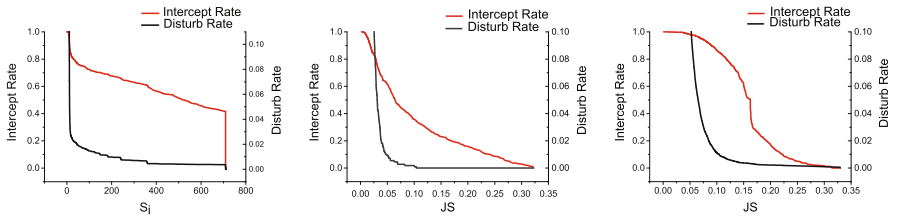

Fig. 4. Trends of different threshold values corresponding to intercept rate and disturb rate.

In the experiment, we randomly selected 1,000 users to replacing their new records with others. Then we use the three methods presented in the Sect. 3 to carry out corresponding identity theft detection. In order to compare the detection performance of our methods, we calculate the experimental results of the IRs, the DRs and the PRs.

4.5 Main Result

We first give the results of the simulation via the three methods we presented in Sect. 3 for two real-life datasets of Foursquare and Yelp. Figure 5 shows the results of three approaches for identity theft detection. The detection method based on check-in has the lowest AUC. The reason is that the data sparsity leads to the low recognition degree of the user spatial distribution. The AUC of DoF detection is the best result, which indicates that social preferences are suitable to reflect user behavior characteristics in MSNs.

As illustrated by the curves in Fig. 5, the effect of identity theft detection is reliable via the above three methods, because the AUC is very large which can reach more than 0.962. However, it is worth emphasizing that the result is obtained on the identifiable users of each dimension. This means that these methods are only feasible for I_{DoC}, I_{DoT}, I_{DoF}, respectively. In fact, behavior data are very sparse for a particular user on each dimension in MSNs. We conduct further experiments to propose an effective method which is detection via their blended behavior.

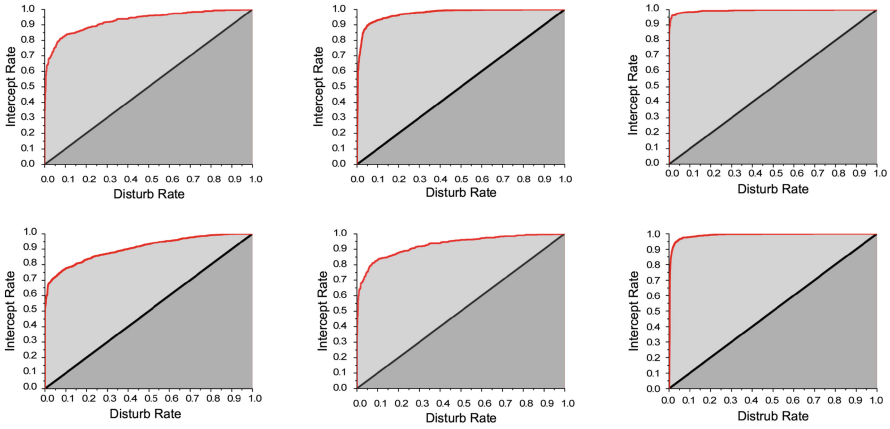

Fig. 5. The ROC curves of three methods on two data set.

We use the blended behavior to perform identity theft detection for all users. We first verify the complementary effect by the fusion between any two dimensions, such as *dimension of check-in and tips* (DoCT). Then, we combine three dimensions together to observe the effect of detection. Table 3 shows the results of seven experiments on two datasets. We get the following conclusions: The first three groups of experiments are the respective results of the above three methods. It is shown that the effect is relatively poor when we consider all users. The *mean intercept rate* (MIR) can only reach 0.518. Secondly, we can observe that when we combine any two dimensions, the performance is significantly improved and the MIR can reach 0.781, while the intercept performance is improved by

50.7%. Thirdly, when we use three dimensions fusion, the IR has reached more than 0.936, which is usually a satisfactory detection performance. At the same time, we also guarantee an acceptable DR at 0.0171. Finally, while the I_{DoT} only accounts for about 4.5%, we can still achieve great detection performance via multi-dimensional fusion due to the complementary effect of blended behavior.

In addition, we also do a similar experiment on the Yelp data, and obtain a similar conclusion. That means the complementary effect of blended behavior can achieve good performance for identity theft detection in MSNs.

5 Literature Review

Identity theft is an ever-present and growing issue in society, where almost all aspects of our lives are digital [5].

Traditional Identification Methods. Studies have shown that the traditional password-based (user-password) authentication technology is still widely used. However the password is easy to leak, easy to forget and easy to copy [6]. Common biometrics usually include fingerprint recognition, face recognition, iris recognition, speech recognition. However, these biometric technology needs to be equipped with high-cost hardware devices which makes the application inconvenient and difficult to popularize. To overcome the drawbacks of the above methods, researchers found that after a user login system, they would produce a series of behaviors and these behaviors include the characteristics of each user. By modeling these behaviors, we can derive behavioral characteristics that uniquely identify the user. Identification based on user behavior came into play [7].

Table 3. Main results of seven identity theft simulations

Model	I_ξ	IR	DR	PR	MIR	MDR	MPR
DoC	18624	0.433	0.008	0.707	0.518	0.006	0.817
DoT	1602	0.336	0.001	0.943			
DoF	17472	0.785	0.008	0.807			
DoCT	18624	0.567	0.009	0.740	0.781	0.012	0.753
DoTF	17681	0.864	0.010	0.806			
DoCF	23525	0.911	0.016	0.713			
DoCTF	23537	0.936	0.017	0.708	0.936	0.017	0.708
DoC	29885	0.226	0.007	0.440	0.396	0.003	0.758
DoT	1912	0.470	0.001	0.977			
DoF	18484	0.492	0.002	0.859			
DoCT	29885	0.648	0.007	0.684	0.681	0.006	0.739
DoTF	19339	0.713	0.002	0.886			
DoCF	42130	0.683	0.009	0.649			
DoCTF	42137	0.872	0.010	0.696	0.872	0.010	0.696

Behavior-Based Identity Analysis. Biometric keystroke recognition technology [9] is through the inherent characteristics of human keystrokes (e.g., keystroke delay and power) for identification, which not only solves the traditional insecurity based on password, but also has the advantages of low cost and high flexibility compared with other biometrics. However, these identification methods usually depend on the specific devices. Once the device is replaced, it takes a long time to retrain. Many researches studied a group or individual behavior features and leverage them to provide a better service [8]. Some of them focused on the spatial-temporal patterns. Cho et al. [2] studied the relation between human geographic movement, its temporal dynamics, and the ties of the social network. Lichman et al. [3] focused on the problem of developing accurate individual-level models of spatial location based on geolocated event data. As one of the most classical probabilistic topic model, LDA is wildly used in modeling text collections (e.g., news articles, research papers and blogs). Yuan et al. [11] propose a novel topic model called SILDA (LDA with Social Interest). Li et al. [10] proposed a new topic model for short texts, named GPU-DMM, which is designed to leverage the general word semantic relatedness knowledge during the topic inference process, to tackle the data sparsity issue. In this paper, we focus on the method for identity theft detection by analyzing user behavior in MSNs.

6 Conclusion

We studied the effectiveness of a promising technique for identity theft detection in mobile social networks, i.e., the method based on user behavioral analysis. To model the user blended behavior in multiple dimensions, we proposed three suitable models in each dimension, such as the user spatial distributions model, user post interests model and user social preferences model. To achieve better detection performance, we proposed a method through multiple dimensions fusion. As a result, we proved the existence of complementary effect of blended behavioral for identity theft detection in MSNs.

Acknowledgments. The research of authors is partially supported by the National Natural Science Foundation of China (NSFC) under Grants 61571331, Shuguang Program from Shanghai Education Development Foundation under Grant 14SG20, Fok Ying-Tong Education Foundation for Young Teachers in the Higher Education Institutions of China under Grant 151066, and the Shanghai Science and Technology Innovation Action Plan Project under Grant 16511100901.

References

1. De Montjoye, Y.A., Radaelli, L., Singh, V.K., et al.: Unique in the shopping mall: On the reidentifiability of credit card metadata. Science 347(6221), 536–539 (2015)
2. Bao, J., Zheng, Y., Mokbel, M.F.: Location-based and preference-aware recommendation using sparse geo-social networking data. In: International Conference on Advances in Geographic Information Systems, pp. 199–208 (2012)

3. Lichman, M., Smyth, P.: Modeling human location data with mixtures of kernel densities. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 35–44 (2014)
4. Zhang, J., Chow, C.: CRATS: an lda-based model for jointly mining latent communities, regions, activities, topics, and sentiments from geosocial network data. IEEE Trans. Knowl. Data Eng. **28**(11), 2895–2909 (2016)
5. Zhou, X., Liang, X., Zhang, H., Ma, Y.: Cross-platform identification of anonymous identical users in multiple social media networks. IEEE Trans. Knowl. Data Eng. **28**(2), 411–424 (2016)
6. Daz-Santiago, S., Rodrguez-Henrquez, L.M., Chakraborty, D.: A cryptographic study of tokenization systems. Int. J. Inf. Secur. **15**(4), 413–432 (2016)
7. Naini, F.M., Unnikrishnan, J., Thiran, P., Vetterli, M.: Where you are is who you are: User identification by matching statistics. IEEE Trans. Inf. Forensics Secur. **11**(2), 358–372 (2016)
8. Wang, C., Zhou, J., Yang, B.: From footprint to friendship: modeling user followership in mobile social networks from check-in data. In: SIGIR 2017, pp. 825–828 (2017)
9. Kambourakis, G., Damopoulos, D., Papamartzivanos, D., Pavlidakis, E.: Introducing touchstroke: keystroke based authentication system for smartphones. Secur. Commun. Networks **9**(6), 542–554 (2016)
10. Zuo, Y., Wu, J., Zhang, H., Lin, H., Wang, F., Xu, K., Xiong, H.: Topic modeling of short texts: a pseudo-document view. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August, pp. 2105–2114 (2016)
11. He, Y., Wang, C., Jiang, C.: Mining coherent topics with pre-learned interest knowledge in Twitter. IEEE Access **5**, 10515–10525 (2017)