User Sensitive Data Identification Method Based on Constraint Gaussian Mixture-Probability Hypothesis Density Filter

Zhengqiu Lu, Shengjun Xue, Chunliang Zhou and Quanping Hua, Defa Hu, Weijin Jiang

***Abstract*—**In order to identify the sensitive data of users in Internet, a sensitive data identification method is proposed by weight constraint Gaussian Mixture-Probability Hypothesis Density (GM-PHD) filter and Restricted Boltzmann Machines (RBM) in this thesis. At first, the data is normalized with weight constraint in this method, and the random network is formed by the definition of the collected characteristic simulation energy function of RBM. Then, the sensitive feature weight of sensitive data is generated in GM-PHD filter. Finally, the simulation experiments are conducted to study this method performance compared with GM-PGD filter, Gaussian filter by MATLAB, including filtering and tracking performance, relevancy degree, sensitive words weight, cluster mapping and high frequency approximation. The results show that, compared with other methods, this method has better performance.

***Keywords*—**sensitive data, weight constraint, Gaussian mixture-probability hypothesis density, restricted Boltzmann machine.

I. INTRODUCTION

WITH the development of science and technology, people’s dependency on the Internet intensifies. Greater attention has been paid to sensitive data with the popularity of applications growing in modern life [1-4]. The common collection of sensitive data involves Oracle Database, Android, cloud environment [5-7] etc. Since the dataflow generated from big data is dynamic, it therefore is influenced by the actual uses and the network environment. As a result, some of the sensitive data cannot be identified effectively when data in large-scale network integration exchange.

The fact that generated data flow can be clustered in data transmission, and some applications have the behavior that notify cluster data flow on their own initiative, some research indicates that the disclosure of sensitive data occurs in the course of initiative notification of cluster information. So, to identify sensitive data effectively is of great significance. As present, there are two major ways of sensitive data identification: data dictionary matching and artificial identification. To prevent the loss in economy and reputation due to the disclosure of sensitive data, some sensitive data can be secured by secret key encryption or another is setting up protection barrier by popularity of cloud computing [8]. Among which the main protective method is to use labels for sensitive data identification in numerous data. Nowadays smart phones are the important collection locations of sensitive data, and some of the Android malware can associate one and another automatically [9]. Literature [10] puts forward an Android malware detection method based on permission sequential pattern mining algorithm, it designs the mining algorithm to permission sequential detection for malware, and warns sensitive information, which could be produced when using malware. However, this method lacks accuracy because the permission mode can be applied in normal applications. Sensitive data plays a significant role in other aspects information, and the database normally protects sensitive data with encryption algorithm, for example, using transparent data to encrypt [11] the sensitive data in Oracle database. However, the access control depends on the authorization of external functions, yet it lacks pertinence identification.

So, this thesis puts forward a method to identify sensitive data based on weight constraint GM-PHD [12-16] filter and RBM [17-19]. It is primarily built on the random Neutral network model based on probability, and which is normalized with weight constraint. And finally, it can extract the features of sensitive data and the structure of belief network effectively. Meanwhile, the successful detection rate of the sensitive data in stimulation neutral network can be improved by calculating the probability of the sensitive words which occur frequently and maximizing the eligible sample probability.

II. SENSITIVE DATA FEATURE MODEL

Sensitive data occur frequently in online applications, and in general, a malware involving sensitive data will generate the cooperation between several permission frequent itemsets. In addition, association rules and cluster mapping are formed

This work was supported in part by Foundation of Zhejiang Educational Committee for contract (Y201738610), and National Natural Science Foundation of China (41275116, 61202464, 61472136 and 61772196).

Zhengqiu Lu is with the Department of Information & Media. Zhejiang Fashion Institute of Technology, Ningbo 315175, Zhejiang, China (corresponding author; e-mail: 459246322@qq.com).

Shengjun Xue is with the Department of Computer & Software, Nanjing University of Information Science & Technology, Nanjing 210044, Jiangsu, China.

Chunliang Zhou is with the Department of Information & Engineering, Dahongying University, Ningbo 315175, Zhejiang, China.

Quanping Hua is with the Department of Information & Media. Zhejiang Fashion Institute of Technology, Ningbo 315175, Zhejiang, China.

Defa Hu is with the Computer and Information Engineering, Hunan University of Commerce, Changsha 410205, Hunan, China

Weijin Jiang is with the Computer and Information Engineering, Hunan University of Commerce, Changsha 410205, Hunan, China

based on the permission feature when sensitive data is transmitted among applications. Here the load is *a*, permission scheduling result is *b*, access data is ω. There are sensitive data interaction between process *i* and *j* which is generated from dataset *u*, and they calculate the scheduling *b* with static method.

$$a = \sum_{i=1}^{n} \omega_i (1 - \mu_i) \tag{1}$$

$$b = k_1 \sum_{i} \omega_i \mu_i + k_2 \sum_{i,j} d_{i,j} \left| \mu_i - \mu_j \right| \tag{2}$$

Here, k_1 and k_2 are correlation weight value respectively, *i, j* = 1, ..., *n*. This model is conducted in the *U* domain, and *T, V* represent the interval range for time and speed. The associate rules scheduling model φ is built on the basic variations mentioned above.

$$\varphi = k_1 \sum_{t\in T} \sum_{v \in V} C_V \left(1 + y_{t,v}\right) + k_2 \sum_{f_i \in MF} \sum_{u,z\in U} \varpi_i d_{u,z} \left|\mu_u^i - \mu_z^i\right| \tag{3}$$

The result of the data statistics indicates that sensitive data usually contains sensitive words, and sensitive data is constantly disclosed without the users being informed. Therefore, the features is extracted and recognized from the sensitive probabilities *m*, *S* and *w* in sensitive data *x* in this thesis, and to conduct the frequency of sensitive words occurring in sensitive data. Assuming that after the interact of HTTP protocol and the server, the probability of the sensitive words occurring in the number of *N* data packets is *f*, the sensitive words weight calculation model ς is proposed here:

$$\varsigma(f) = \sum_{N} U_T^{-k_1 m^T (T_N - \tilde{T}_N)} w^0 (x_i, \tilde{x}_i, S_i) + k_2 v_i w^T (T_N, \tilde{T}_N) \tag{4}$$

Here, *U* is the high degree of approximation:

$$U_{k|k-1}(\mathrm{x}) = \sum_{i=1}^{f_{k|k-1}} w_{k|k-1}^{(i)} N\left(x; m_{k|k-1}^{(i)}, S_{k|k-1}^{(i)}\right) \tag{5}$$

The occurrence probability of sensitive words is calculated using sensitive words weight in this thesis. Further study of the cluster behavior among sensitive data is conducted in order to identify sensitive data accurately. In this thesis, the number of sensitive sample sets is *n*, *r* is target sample, y_i is the feature signature obtained by collecting and recognizing sensitive data. And the mapping cluster model ξ combined with based on Gaussian Mixture mode is here.

$$|\xi(r)| = \left((y_i(r,1) - y_i(r,1)^2) + \cdots + \left(y_i(r,n) - y_i(r,n)\right)^2 \right)^{\frac{1}{2}} \tag{6}$$

$$y_i = \left[-x_i lg \tilde{x}_i - x_i \lg(\widetilde{x_i})\right] \left\| T_N - \tilde{T}_N \right\|^2 \tag{7}$$

III. IDENTIFICATION METHOD

In order to identify sensitive data effectively and accurately, this thesis does some research based on weight constraint GM-PHD filter and RBM. It can track the unknown objectives by GM-PHD filter without particle collection and filter cluster, and keep the information consistent using Restricted Boltzmann Machines, while the sensitive words having density filter. Considering the fact that self-adaptive GM-PHD is unable to track and cross objectives on a one-to-one basis, this thesis therefore aims to improve GM-PHD filter combined with constraint weight.

In order to achieve the sensitive feature of sensitive data, and filter the interference factors unrelated to objective, this thesis gets the feature by RBM. After the target environment is determined, data is grouped according to its sensitive feature. The RBM proposed in this thesis includes foreground hidden layer and background hidden layer, aiming at the environment factor *v* and time factor *T* of target signal. The feature identification mode λ is built on the basis above:

$$\lambda = \sigma(v_i (b_i^1 + b_i^2 + \sum_{j=1}^{n} T_{ij}^1 \left(1 + h_j^1\right) + \sum_{j=1}^{n} T_{ij}^2 \left(1 + h_j^2\right))) \tag{8}$$

Here, σ represents weight coefficient, λ1 and λ2 are the feature distribution functions based on foreground and background layers.

$$\lambda_1 = \sigma\left(c_j + \sum_{j=1}^{m} v_i T_{ij} \right) \tag{9}$$

$$\lambda_2 = \sigma\left(b_j + \sum_{j=1}^{n} h_i T_{ij} \right) \tag{10}$$

In order to identify the sensitive data in target data effectively, this thesis firstly processes the data in Restricted Boltzmann Machines, and calculates the sensitive word weight and high-frequency approximation of sensitive data, and then describes the data clustering behavior according to the cluster mapping property of sensitive data. The algorithm is stated as follow in details:

(1) Set up the environment. Build a special topology structure of RBM, including visible neuron layer, environment neuron layer, background neuron layer and hidden neuron layer. In the visible neuron layer, we seize and determine two target signals ω and θ, track and identify the tracking signal by filter, and then input the identified data into visible neuron layer.

(2) Initialize the parameters. Define dynamic changing model and sensor measurement model according to target objectives x and y. *N*(*x; m; p*) represents Gaussian distribution, in which *m* represents the mean value and *p* represents covariance, ω represents target weight, Q_{k-1} represents the defined covariance in data transmission and F_{K-1} represents the transfer matrix of target state. The predictive function *L* for sensitive data initialization is defined on the basis of Gaussian

distribution sensor model *f*.

$$f_{k|k-1}(x,y) = N(x; F_{K-1}; Q_{k-1}) \qquad (11)$$

$$L(\mathrm{x},\mathrm{y}) = \sum_{i=1}^{k-1} \omega_{k-1}^{(i)}\, N(x; m_{k-1}^{(i)}, P_{k-1}^{(i)}) N(y; m_{k-1}^{(i)}, P_{k-1}^{(i)})(12)$$

(3) Calculate the result of permission scheduling based on formula (1-3) in the neuron. Then predict the sensitivity of the data according to formula (11) and (12). After that, calculate the relevance of the data and input the result into environment neuron layer by formula (4) and (5).

(4) Calculate the model based on state and data. If the target does not coincide with the one-to-one hypothesis, the weight of the target should be altered by standardizing the behavior. The GM-PHD filter algorithm of constraint weight primarily can divided into two steps including predicting and updating. If the prediction intensity function *v* which identify *x* and *y* can describe the Gaussian Mixture form.

$$v_{k|k-1}(x,y) = P_{s,k} \sum_{j=1}^{k-1} \omega_{k-1}^{(j)} N\left(x; m_{S,k|k-1}^{(j)}, P_{S,k|k-1}^{(j)}\right) \qquad (13)$$

$$m_{S,k|k-1}^{(j)} = F_{k-1} m_{k-1}^{(j)} \qquad (14)$$

$$P_{S,k|k-1}^{(j)} = Q_{k-1} + F_{k-1} P_{k-1}^{(j)} F_{k-1}^{T} \qquad (15)$$

By restricting the weight of average target number, the average root-mean-square error of target number and the distance of OSPA, the limitation that GM-PHD filter cannot process data on the one-to-one basis. Assuming that there are two limited signal sets X={x_1,x_2,...,x_m} and Y={y_1,y_2,...,y_n} , whose parameters are valued as m≤n, and d_c indicates the stage distance between two test signal sources. *P* and *c* are the false-alarm and undetected parameters, and they have to fulfill the following formula to accord with one-to-one tracking.

$$\tau(\mathrm{X},\mathrm{Y}) = (\frac{1}{n}(min \sum_{i=1}^{m} d_c\left(x_i, y_{\alpha(i)}\right)^P + c^P(n-m)))^{\frac{1}{P}} \quad (16)$$

(5) Calculate the environment factor in neuron layer combing information data. An approximation of high frequency can be obtained according to formula (13) and its permission can be restricted with formula (16) to enable the filter to track the target on the one-to one basis.

(6) Acquire the sensitive data features. Input the acquired information into the background neuron layer and use formula (8) to obtain the sensitive words features and then input the data into the next neuron.

(7) Analyze the cluster mapping. Analyze the cluster mapping in the hidden neuron layer, and acquire the feature data of each neuron layer using formula (16) including sensitivity characteristics of sensitive data, high frequency approximation, sensitive words permission and cluster mapping under that environment.

(8) The algorithm is finished.

IV. SIMULATION EXPERIMENTS

In this thesis, the stimulation experiment is conducted by the proposed identification method based on MATLAB. We seize 10000 data packets by terminal accessing application, including cluster data packets, dispersion data packets and ordinary data packets, among which cluster data is assumed to consist of sensitive data packets and ordinary data packets. We choose several data packets randomly to conduct the stimulation test and compare ordinary GM-PHD filter with Gaussian filter. Since the weight of GM-PHD filter for the tracked target is constrained, the tracking performance of the data is enhanced as the noise for each frame in the data packet which is extracted, and the result illustrate in Fig 1.

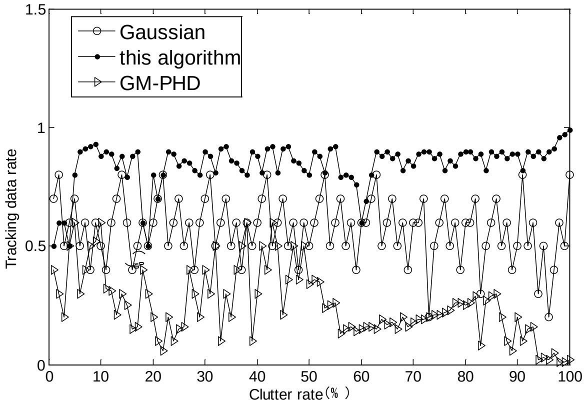

Fig. 1 Comparison of Filtering and Tracking Performance

Secondly, in order to verify the one-to-one tracking efficiency of the algorithm for the target, the relevancy degree of the three algorithms stated above is compared. The target data transmission is normalized and the range of relevancy is specified from 0 to 100. The relevancy degree of distributed data is calculated with three algorithms, as illustrated in Fig 2, from which it is clear that the relevancy degree between sensitive data and data will increase as the data interact more frequently, and when the data stream peaks at the point of unrestricted weight algorithm, the relevancy degree of data can be calculated effectively with the algorithm proposed in this thesis.

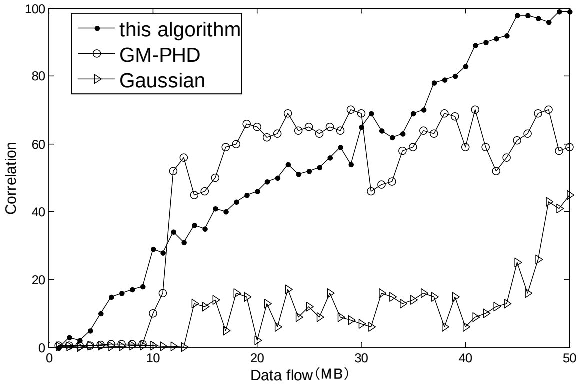

Fig. 2 Comparison of Relevancy Degree

Meanwhile, the weight of output sensitive words is taken for an important index to weigh the sensitive data. As a result, the disclosure of sensitive data can be monitored according to its weight. As Fig 3 indicates, the values of the sensitive words weight from three algorithms are compared, and the range of the value is limited from 0 to 60. It is clear from Fig 3 that most range of the weight value can be monitored with the algorithm proposed in this thesis while the other two algorithms cover a smaller range.

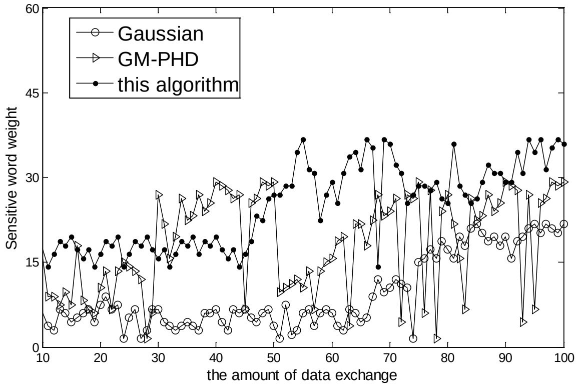

Fig. 3 Comparison of Sensitive Words Weight

Finally, the comparisons of high frequency approximation and cluster mapping from the three algorithms are provided in Fig 4 and Fig 5 respectively. It can be concluded that as the degree of cluster increases, the features of sensitive data can be reflected from the cluster mapping generated by clustering of sensitive data. As Fig 4 illustrates, GM-PGD filter, Gaussian filter and the algorithm proposed in this article have similar capacities to describe the approximation of the high frequency when the data volume is relatively small.

However, when the data volume is large, the algorithm proposed in this thesis achieves a better performance in identifying the approximation of high frequency, which reflects that this algorithm has more effective in data tracking and better adaptively. According to Fig 5, the algorithm proposed in this thesis has superior cluster mapping property and can provide better reliability in data identification with the same permission relevancy.

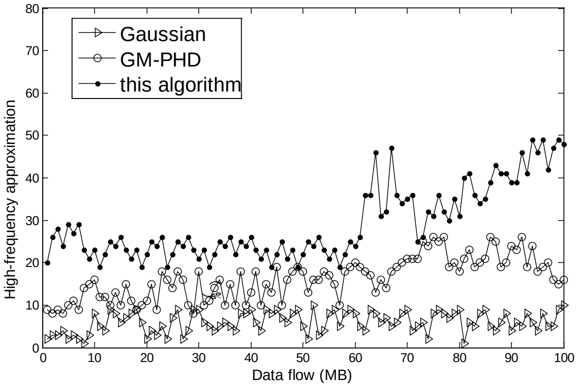

Fig.4 Comparison of High Frequency Approximation

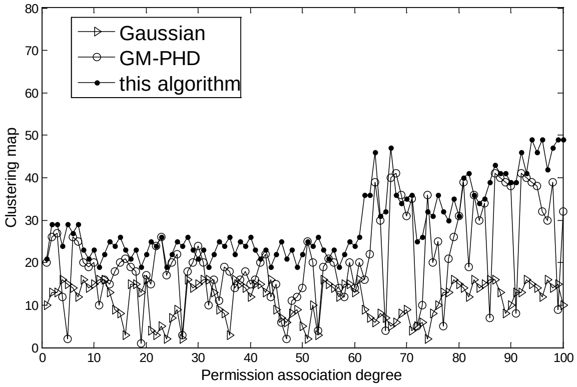

Fig. 5 Comparison of Cluster Mapping

V. CONCLUSION

The disclosure of sensitive data on the Internet is primarily due to the initiative behavior of sending cluster information, and these sensitive data has high frequency approximation and cluster mapping. In order to prevent the disclosure of sensitive data based on sensitive words, a sensitive data identification method based on weight constraint GM-PHD filter and RBM is put forward in this thesis. It normalizes the data by weight constraint relationship, and forms the random network by definition of RBM's collection feature simulation energy function, and then generates the data sensitive feature weight in in GM-PHD filter. Finally, in order to verify the method performance, we compare the performance of common GM-PHD filter and Gaussian filter by the simulation experiments with MATLAB. The results show that the method proposed in this thesis has higher adaptability.

REFERENCES

[1] WU Jingzheng,WU Yanjun,WU Zhifei etc. “An Android privacy leakage malicious application detection approach based on directed information flow”,*Journal of University of Chinese Academy of Sciences*, vol. 32, no.6, pp. 807-815,2015.

[2] Liu Aihua,Chen Jun,Xie Fang. “Research and Improvement of Sensitive Data Encryption Algorithm in Database”,*Journal of Nanjing Normal University (Engineering and Technology Edition),* vol. 12, no.9, pp. 68-71,2012.

[3] XU Jiang-Feng Zhuang Hai-Yan,YANG You.“The Analysis of Encryption Technology of Oracle Database”, *Computer Science*, vol. 33, no.1, pp. 134-136,2006.

[4] FAN Jing,SHEN Jie,XIONG Li-rong. “Scheduling Data Sensitive Workflow in Hybrid Cloud”,*Computer Science*, vol. 42, no.11, pp. 400-405,2015.

[5] CHEN Wei,WANG Yi,QIN Zhi-guang etc. “Research on Timed Access of Sensitive Data Based on Dual Encryption”,*Journal of University of Electronic Science and Technology of China*, vol. 46, no. 3, pp. 588-593,2017.

[6] LI Hai-Feng,ZHANG Ning,ZHU Jian Ming etc.“Frequent Itemset Mining over Time-Sensitive Streams”,*Chinese Journal of Computers*, vol. 35, no. 11, pp. 2283-2293,2012.

[7] HAN Xinhui,WANG Dongqi,CHEN Zhaofeng etc. “Study of A Protection Method of Sensitive Data of Web Servers in The Cloud”,*Journal of Tsinghua University(Science and Technology)*, vol. 56, no.1, pp. 51-57, 2016.

[8] Crampton J, Koponen T. “Delegation in role-based access control”, *International Journal of Information Security*, vol. 7, no.2, pp. 123-136, 2008.

[9] WANG Zhiqiang,ZHANG Yuqing,LIU Qixu."Algorithm to detect Android malicious behaviors",*Journal of Xidian University*, vol. 42, no.3, pp. 8-14,2015.

[10] YANG Huan,ZHANG Yuqing,HU Yupu etc."Android malware detection method based on permission sequential pattern mining algorithm",*Journal on Communications*, vol. 34, no.1, pp. 106-115,2013.

[11] LIU Zongxiang,XIE Weixin,WANG Pin etc. "A Gaussian Mixture PHD Filter with the Capability of Information Hold",*Acta Electronica Sinica*, vol. 41, no.8, pp. 1603-1608,2013.

[12] LI Haifeng,ZHANG Ning,ZHU Jianming etc."Frequent Itemset Mining over Time-Sensitive Streams",*Chinese Journal of Computers*, vol. 35, no.11, pp. 2283-2293, 2012.

[13] CHEN Jinguang,QIN Xiaoshan,MA Lili. "Fast GM-PHD for Multi-target Tracking",*Computer Science*, vol. 43, no.3, pp. 316-321,2016.

[14] CANG Yan,CHEN Di,BI Xiaojun."Application of adaptive GM-PHD filters to multi-target tracking",*Journal of Harbin Engineering University*, vol. 36, no. 11, pp. 1526-1531, 2015.

[15] ZHANG Yifeng."Multi-target Tracking Method Based on GM-PHD Filtering with Weight Constraint",*Computer Engineering*, vol. 43, no. 3, pp. 282-288,2017.

[16] LIU Liangkun,ZHANG Dakun."Improved Algorithm for Finding Weight-constrained Maximum-density Path",*Computer Science*, vol. 41, no. 8, pp. 122-124, 2014.

[17] HE Jieyue,MA Bei."Based on Real-Valued Conditional Restricted Botzmann Machine and Social Network for Collaborative Filtering", *Chinese Journal of Computers*, vol. 39, no. 7, pp. 183-194,2016.

[18] YANG Jie,SUN Yadong,ZHANG Liangjun etc."Weakly Supervised Learning with Denoising Restricted Boltzmann Machines for Extracting Features", *Acta Electronica Sinica*, vol. 42, no. 12, pp. 2365-2370,2014.

[19] WANG Yingxue,ZHANG Shenghui,YU Yingying etc."Speech Bandwidth Extension Based on Restricted Boltzmann Machines", *Journal of Electronics & Information Technology*, vol. 38, no. 7, pp. 1717-1723, 2016.

Zhengqiu Lu was born on Nov. 26, 1982**.** He received the Master degree in computer application technology from Wuhan University of Technology of china. Currently, he is a researcher (lecturer) at Zhejiang Fashion Institute of Technology, China. His major research interests include wireless network and data processing. He has published many papers in related journals and conference.

Shengjun Xue was born on Dec. 5, 1956**.** He received the PhD degree in computer application technology from Wuhan University of Technology of china and received the Postdoctoral degree in Purdue University. Currently, he is a researcher (professor) at Nanjing University of Information Science & Technology, China. His major research interests include computer network, and cloud computing. He has published many papers in related journals and conference. In addition, he was a chair of International Conferences such as RSETE,CSSS in past five years.

Chunliang Zhou was born on Apr. Oct, 1982**.** He received the Master degree in computer application technology from Ningbo Institute of Technology, Zhejiang University of China. Currently, he is a researcher (lecturer) at Dahongying University of, China. His major research interests include wireless sensor network. He has published many papers in related journals and conference.

Quanping Hua was born on Jan. 13, 1968**.** He received the master degree in computer science from Southeast University of china. Currently, he is a researcher (professor) at Zhejiang Fashion Institute of Technology, China. His major research interests include data mining. He has published many papers in related journals and conference.