Multiple methods for wechat identification

Chunwei Tian[1, a], Qi Zhang[2,b] and Guanglu Sun[1,c]

[1] School of Computer Science and Technology, Harbin University of Science and Technology, 150080, China

[2] School of Economics and Management, Northeast Agriculture University, 150030, China

[a]chunweitian@163.com, [b]leaderman2145337@163.com, [c]guanglu_sun@163.com

Keywords: Traffic classification; Wechat identification; Malware detection; text classification; image classification

Abstract. Wechat is a popular social platform developed in 2011 by Tencent. In this paper wechat traffic is analyzed by a novel hybrid method, which combines statistical method, payload-based method, SVM, CRC and deep learning. Firstly, the statistical method is utilized to extract features from wechat packets header, which can classify different wechat applications and functions, e.g. texts, images and voice, and so on. Secondly, payload-based method is used to identify the traffic, which is corresponding to the above functions and application protocols. Thirdly, SVM is applied to categorize the texts based on their attributes. CRC method is used to classify the images, which effectively protects the user's privacy. Finally, deep learning is presented to extract features of wechat app in order to check the malicious software. Experimental results show that, the proposed method has high accuracy for wechat traffic. It not only identifies wechat app, but also detects the specific functions of app. It even discriminates texts, images, voice and malicious software effectively.

Introduction

Network technology developed fast in the past 20 years. There are lots of innovative applications such as file sharing, instant messaging and streaming media transmission which have been widely employed in human's life [1]. Social network platform is one of the concerned hotspots [2]. Traditional traffic classification methods mainly include port-based method, payload-based method, behavior-based method and machine learning method. Karagiannis et al. [3] proposed port-based method to obtain single traffic, which is effective to test the performance of other methods. However, it has low identification rate for traffic. Sen et al. [4] proposed application protocol features method to recognize 5 p2p protocols, which obtains better performance for traffic identification. However, it has high complexity of algorithm. Hwang et al. [5] proposed support vector machine (SVM) method to recognize traffic, which may be effective for traffic classification.

Multiple methods are used to classify traffic, which have obtained better performance than single method for traffic identification [6]. The combination of multiple methods is first proposed to analyze wechat traffic. The main steps are as follows. Firstly, statistic method is exploited to extract features and identify wechat specific functions, e.g. texts, images and voice. Secondly, payload-based method is employed to identify protocols of special functions. Thirdly, SVM method is exploited to classify text. Then, collaborative representation classification (CRC) is utilized to classify the images of wechat app. Finally, deep learning method is used to extract features of wechat app and detect if the wechat app includes malicious software. This novel method has the following merits. First of all, this novel method can recognize wechat app and its specific functions. Next, it can identify partial malicious software. Finally, the novel method is flexible and easy to implement, so it has very good prospect for practical applications.

The remainder of this paper is organized as follows. Section 2 introduces the proposed novel method. Section 3 shows results of some experiments. Section 4 provides our conclusion.

The proposed method

A. The algorithm obtains the features for wechat app identification

To identify wechat app, based on statistical method is exploited to obtain relational features and identify wechat app, which obtains good performance. We introduce the implementation of based on statistical machine learning method for wechat identification as follows. First, wireshark and GT tools are chosen to capture pure wechat traffic. Next, we obtain payload of transport layer of wechat packet. Then, we count the probability, which presents character of the same position and the same byte. When probability of above character is over 0.95, we regard them as a feature of wechat identification. Finally, we use obtained features to identify non-single wechat app and use GT tool to calculate the identification rate of wechat app. This proposed method extract them features, we also call them header format of wechat packet as follows. The header format of wechat packet has 16 bytes: the length of wechat packet (4 bytes), the length of header (2 bytes), version (2 bytes), special function number (4 bytes) and serial number (4 bytes). Different function number stands for different function. If the function number is 0xed, it would represent text message. If the function number is 0x13, it would represent voice message. If the function number is 0x09, it would represent image message. If the function number is 0x62, it would represent circle of friends. If the function number is 0x06, it would represent heartbeat packet. So the proposed method identifies not only the wechat traffic, but also can distinguish specific function.

B. The payload-based method identifies protocols of special function

The payload-based method has high accuracy for wechat identification. We use the first 70 bytes of wechat payload to identify protocol of special functions. We utilize deep packet inspection (DPI) to identify wechat functions as follows. First step obtains first 70 bytes of special wechat functions. Second step uses payload-based method to identify protocol of special functions. The experiments show the protocols of voice message and image message are http.

C. The SVM method classifies text of wechat app

Support vector machine (SVM) is exploited to classify text of wechat app in this paper. SVM can use hyper-plane $w \cdot x + b = 0$ to classify samples, especially $x=(x_1,x_2,x_3,\ldots,x_n)$ is vector of sample feature and hyper-plane is decided by normal vector w and intercept b. Sample space is divided into two areas: positive sample and negative sample. Normal vector w points the direction of positive sample.

As fig.1 shows, H is hyper-plane, A, B, C and D become support vector. Support vector A and B obtain interval boundary H_1: $w \cdot x + b = 1$. Support vector C and D obtain interval boundary H_2: $w \cdot x + b = -1$.We define that distance of the boundary intervals is called as margin, the value of the interval is determined by the normal vector w of hyper-plane. The value of the interval is $\frac{2}{\| w \|}$. When the training samples are separable, numerous hyper-planes can separate positive and negative samples, so SVM ensures that make interval maximum and hyper-plane has unique solution. The solution of SVM is Ref. [7].

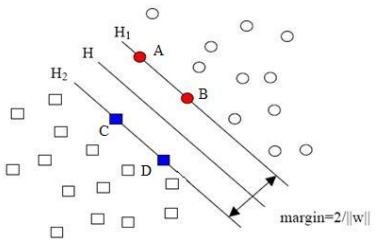

Fig1. Linear separable support vector machine maximum margin learning method

The SVM classification method has the following phases. First phase obtains first 64 bytes of text message. Second phase uses SVM to classify text message. Our proposed method has good performance for text classification. The results of classification have two categories: bad text and good text.

D. CRC method classifies images

Collaborative representation classification has good performance on image classification. We think that all the training samples of each subject are over-enough in the discussion of sparse. However, face image is small sample and it is often not enough. We regard face images of other subjects as training samples of the same test sample when samples are not enough. We put test sample use collaborative representation to represent and use collaborative representation classifier to classify images [8].

The implementations of CRC method for image identification have the following steps. First, we obtain images from wechat by manual settings. Second, we divide images into two parts: training dataset and test dataset. Finally, we use CRC to classify images. Our method is easy and simple to implement, which has good prospect for real applications. Our method has high accuracy for image classification, which improves privacy.

E. The deep learning method identifies malware

We use statistical analysis and dynamic analysis to extract 200 features from each android app, and then we employ deep learning technology to classify the malware from normal apps [9]. Finally, we make sure whether wechat app has malware plug-ins.

EXPERIMENTS AND RESULTS

We use two parts to show the performance of this novel method. First part is used to show the identification rate of wechat. The Second part is exploited to illustrate the accuracy of classification images.Three kinds of wechat data are chosen to test the performance of wechat traffic. Meanwhile, we also choose different system to verify the performance of the proposed method for wechat traffic identification. These systems include ios and windows. Three datasets make up with wechat of tcp long connection, non-wechat traffic and mixed traffic (both of wechat traffic and non-wechat traffic).To avoid the chance of obtained features, we employ wechat platform of ios and web wechat of windows to send wechat packets, respectively. The first dataset includes 305 wechat packets (60 flows). The second dataset includes 289 wechat packets (46 flows) and 67 non-wechat packets (46 flows). The third dataset includes 233267 non-wechat packets (21849 flows). The forth dataset includes 157 wechat packets (19 flows). The fifth dataset includes 48 wechat packets (3 flows) and 36 non-wechat packets (8 flows). The final dataset includes 62507 non-wechat packets (2965 flows). Our experiment is illustrated as Table 1.

TABLE 1. IDENTIFICATION RATE OF WECHAT TRAFFIC

	DATASET	identification rate(%)	
		packet	flow
IOS	1th group	97.09	95.86
	2th group	95.38	93.1
	3th group	0	100
WEB	4th group	93.80	93.61
	5th group	99.1	98.1
	6th group	0	100

To fully test the performance, we choose public ORL face dataset to conduct the experiment. As shown in Fig. 2, we choose 5 original images to show ORL.We use CRC to classify images, which obtains good performance. The results of experiment are as Table 2. The results of experiment show that the novel method has good performance.

Fig 1. 5 original images from ORL face dataset

TABLE 2 RATE OF CLASSIFICATION ERRORS(%) FROM THE ORL DATASET

Number of training samples each class	1	2	3
The proposed novel method with collaborative representation	32.65	23.01	18.49

Conclusions

This paper uses the combination of statistical method, payload-based method, SVM, CRC and deep learning to classify wechat traffic. The proposed method not only identifies wechat traffic, but also distinguishes special functions of wechat app and malicious software. It analyzes wechat from different views and levels. So it has important significance for special applications. The good performance of the novel method has been proved in the experiments. The proposed method can be easily implemented.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No.61502123), Natural Science Foundation of Heilongjiang Province of China (No.QC2015084).

References

[1] K. D. Zeilenga, Internet Assigned Numbers Authority (IANA) Considerations for the Lightweight Directory Access Protocol (LDAP).(2002)

[2] Z. Chu, C. Gian and H. Wang. Detecting automation of twitter accounts: Are you a human, bot, or cyborg?, Dependable and Secure Computing, IEEE Transactions on, 9(6): 811-824. (2012)

[3] T. Karagiannis, A. Broido, M. Faloutsos and k. c. claffy. Transport layer identification of P2P traffic, In proceeding of the 4th ACM SIGCOMM conference on Internet measurement, New York, NY, USA, 134-138. (2004)

[4] S. Sen, O. Spatscheck and D. Wang, Accurate, scalable in-network identification of p2p traffic using application signatures, In Proceedings of the 13th international conference on World Wide Web, ACM, 512-521. (2004)

[5] S. Hwang, K. Cho and J. Kim, Traffic classification approach based on support vector machine and statistic signature, Springer Berlin Heidelberg, 332-339. (2013)

[6] T. Karagiannis, K. Papagiannaki and M. Faloutsos, BLINC: Multilevel Traffic Classification in the Dark, In SIGCOMM'05, Philadelphia, USA, 229-239. (2005)

[7] D. Zuev and A. W. Moore, Traffic classification using a statistical approach, Passive and Active Network Measurement, Springer Berlin Heidelberg, 321-324. (2005)

[8] Y. Xu, B. Zhang and Z. Zhong, Multiple representations and sparse representation for image classification, Pattern Recognition Letters, 68: 9-14. (2015)

[9] Z. Yuan, Y. Lu and Z. Wang Droid-sec: Deep learning in android malware detection. ACM SIGCOMM Computer Communication Review. ACM, 44(4): 371-372. (2014)