Behavior Rhythm: A New Model for Behavior Visualization and Its Application in System Security Management

Zhaoli Liu[1], Xiaohong Guan[2,1], (Fellow, IEEE), Shancang Li[3],Tao Qin[1,2], and Chao He[1]
[1]Key Laboratory for Intelligent Networks and Network Security of the Ministry of Education, Xi'an Jiaotong University, Xi'an 710049, China
[2]Shenzhen Research School, Xi'an Jiaotong University, Shenzhen 518057, China
[3]School of Computer Science, University of the West of England, Bristol BS16 1QY, U.K.

Corresponding author: Xiaohong Guan and Tao Qin (e-mail: xhguan@mail.xjtu.edu.cn, qin.tao@mail.xjtu.edu.cn).

This work was supported in part by the National Natural Science Foundation of China under Grant 61772411, Grant 61672026, Grant 61602370, and Grant U1736205, in part by SZSTI under Project JCYJ20170816100819428, and in part by the Fundamental Research Funds for the Central University. The work of Z. Liu was supported by the China Scholarship Council under Grant 201706280180.

ABSTRACT The widespread use of social media, cloud computing and Internet of Things (IoT), generate massive behavior data recorded by system logs, how to utilize these data to improve the stability and security of these systems becomes more and more difficult due to the increasing number of users and amount of data. In this paper, we propose a novel model named Behavior Rhythm (BR) to characterize and visualize the user's behaviors from the massive logs and apply it to the system security management. Based on the BR model, we conduct the clustering analysis to mine the user clusters. Different management and access control policies can be applied to different clusters to improve the management efficiency. Then we apply the NMF method to analyze the BRs and perform abnormal detection, and employ the BR similarity calculation to perform fast potential anomaly tracking. The detection and tracing results can help the administrators to control the threats efficiently. Experimental results based on the datasets collected from the campus network center of Xi'an Jiaotong University verify the accuracy and efficiency of our method in user behavior profiling and security management, which lay a solid foundation for improving system stability and quality of service.

INDEX TERMS System management, Behavior Rhythm, clustering, NMF, anomaly detection and tracing.

I. INTRODUCTION

The development of social media, cloud computing and Internet of things (IoT) greatly benefits our daily lives, however, the security and privacy management in these platforms become more and more difficult due to the increasing number of users and amount of data. The system logs record the runtime information about system operations, which is an important data source for user behavior analysis and system security management. Traditional log based abnormal detection methods use the patterns of known attacks to identify anomalies, thus these methods cannot deal with the new or unknown threats, which are not sufficient to assure the system stability. The increasing scale and complexity of modern systems make the volumes of logs increase significantly, which poses another challenges to the traditional methods and makes them more time consuming. Facing the above challenges, we propose a novel model named Behavior Rhythm (BR), which is constructed based on the features extracted from different kinds of logs. The BR model can be used to visualize the users' behavior characteristics and help the administrator to master the user's macroscopic behavior characteristics and perform efficient management. Based on the BR model, we employ the clustering algorithm and Non-negative Matrix Factorization (NMF) to achieve the goal of system management and abnormal behavior detection with high efficiency.

Firstly, we treat one account using specific IP address as a user, the user's behavior traces are recorded by different kinds of system logs. The logging activities and the commands extracted from the system logs are the most important data sources for system security monitoring. To fully utilize the characteristics of the logging activities and commands used, we first divide the commands into 8 categories based on their functions, such as the commands related to the system configuration, commands related to the file operations, etc. Meanwhile, we divide one day into

24-time windows and each time window represents one hour. Then we construct a model named as BR, which is composed of some small squares decided by the command type and time window, to visualize the user's behavior characteristics in an easily understandable way. We employ the number of dots in the specific square to indicate the frequency of the user using the corresponding command type during the corresponding time window, more dots indicate higher frequency. The dots with different colors are used to identify different types of the logging IP addresses. We simply divide logging IP addresses into three types, including the CERNET (China Education and Research Network) which is mainly used by the students and teachers, the second type is the other IP addresses in China and the third type is the foreign IP addresses. Based on the BR model, the users' behavior characteristics can be visualized in an easily understandable way, which can help the administrators to master the user's macroscopic behavior characteristics and improve the behavior pattern profiling efficiency.

Secondly, we regard each square in the BR as an element in a matrix to quantify the BR, but each element contains three frequency values corresponding to the three types of logging IP addresses, thus the BR can be converted to a 3*8*24 matrix. We analyze the users' behavior characteristics and find that users usually have obvious routine patterns. They login the system and use similar commands during the similar time window, thus the BRs obtained for the same user in different days are similar. The users with similar behaviors also have similar BRs. Based on these findings, we apply DBSCAN [1] clustering method to classify the BRs obtained, the users with similar behavior patterns are classified into the same cluster. To deeply investigate the users' behavior characteristics in different clusters, we propose the Operation and Maintenance Frequency (OMF) to quantify the behavior characteristics by integrating the logging related features, such as the logging time, logging IP address, session lifetime and the number of commands performed. We also employ the Prefixspan [2] algorithm to mine the frequent command sequence used to infer the user's role and operation purpose in specific cluster. Combined with the clustering and analysis results, we can design suitable security and access control policies for the users in different clusters instead of individual user, which can greatly improve the system management efficiency.

Thirdly, we apply the Non-negative Matrix Factorization (NMF) [3] method to extract the hidden behavior patterns from the BRs and perform abnormal detection. Based on NMF, the data matrix can be factorized into two smaller matrices with positive elements, one as the base and the other as the encoding coefficients representing normal or abnormal features, which are used to perform abnormal detection. Based on the analysis of typical abnormal BRs, we propose a fast anomaly tracking method based on similarity calculation. The network administrators can use the BRs of typical anomalies, or even draw a specific BR corresponding to the behavior patterns they are interested in, to trace the BRs similar with the template BR from the actual network. In this way, they can not only track specific type of anomalies, but also trace the potential threats with some special behavior characteristics efficiently, in turn, help them to keep the system under control and improve the service stability.

To bridge the gap between research and practice, we construct a log collection and security management platform in the campus network center of Xi'an Jiaotong University. We introduce Flume [4], HDFS (Hadoop Distributed File System) [5] and Spark [6] technologies into the platform as parallel data processing and reliable data storage mechanism which can help to process massive data more efficiently and stably. We collect system logs from more than fifty actual running network servers and labeled abnormal logs from one target server based on the platform to verify the proposed method. Experimental results show that the proposed method has high detection accuracy and low computational complexity.

The following of this paper is organized as follows. We briefly review the related works in Section 2. Section 3 describes the framework of the proposed methods and the BR model. Section 4 presents the BR clustering method. The abnormal behavior detection using NMF method and the fast anomaly tracking method are presented in Section 5. Datasets used and experimental results are presented in Section 6. Then conclusion and future work follow.

II. RELATED WORKS

User's behavior monitoring and security management have attracted many attentions and been extensively explored in recent years. Most of previous works are mainly focusing on signature-based anomaly detection methods. The authors in [7] focus on using log analysis techniques for collecting technical security metrics from different kinds of security logs, including the IDS (Intrusion Detection Systems) logs, workstation logs etc., and construct a production framework for anomaly detection. The authors in [8] propose a method for detecting malicious executables with minimum priori information using Multi-Label Naive Bayes classification algorithm. However, this kind of algorithms can only achieve good results in detecting known anomalies, they generally do not perform well against the unknown anomalies, which are not sufficient to assure the system stability.

To face the challenge, many researchers began to explore anomaly detection methods using statistical analysis and data mining techniques. The authors in [9] apply statistical log analysis to improve the scalability of behavioral analysis on massive logs, which uses trace sampling and statistical inference techniques to analyze a sample of traces to compute the statistical guarantees for the analysis results. The authors in [10] employ a statistical learning technique to constant linear relationships from console logs, these linear relationships can capture the normal program execution behavior, if a new log breaks certain invariants, then they declare an anomaly occurs during the system execution. In [11], the authors employ the statistical template extraction (STE) and log tensor factorization (LTF) to mine the potential events. STE focuses on extracting primary templates from the logs using a statistical clustering method. LTF aims to build a statistical model that captures spatial-temporal patterns to provide useful insights for the potential events. The authors in [12] employ the nonnegative matrix factorization (NMF) algorithm to capture each host's network behavior patterns from the logs and perform anomaly detection. In [13], the authors analyze the frequent sequences of the logs and build robust temporal user profiles to model the relation between the users' tasks and their temporal properties, then use it to detect abnormal behaviors. The authors in [14-15] extract the correlations between the logs and apply principal component analysis (PCA) to detect anomalies, then construct decision trees to classify the detected anomalies. The authors in [16] propose an unstructured log analysis technique to convert log messages to log keys, and then learn a finite state automaton (FSA) from training log sequences for anomaly detection. However, along with the increasing scale and complexity of modern systems, how to characterize and model the behaviors from the massive data efficiently becomes more and more difficult.

Network traffic is one of the important data sources to mine user behaviors and attracts many researchers' attention in recent years. With the use of encryption technology, many researchers turn to explore behavior patterns from the connection patterns. The authors in [17] use the traffic connection patterns between end hosts to build the users' behavior profiles, and then they use entropy based techniques to detect anomalies. The authors in [18] use the bipartite graphs and one-mode projection graphs to analyze the communication behaviors between end-hosts, then they apply clustering algorithm on the similarity matrices of the projection graphs for detecting anomalous patterns. The authors in [19] analyze the traffic behaviors by classifying network traffic behavior using machine learning methods. The authors in [20] focus on the aggregated traffic behavior analysis under different network prefixes to mine the inherent behavior clusters in large-scale network. However, the traffic behavior analysis methods cannot detect masquerade and some anomalies at the host level.

To detect the masquerade and the anomalies at the host level, anomaly detection techniques have also been developed at user behavior level based on different kinds of data, such as the command sequences, which are closely tied to user behaviors. The authors in [21] attempt to detect masquerades by building normal user behavioral models using truncated command sequences, they compared several statistical techniques including uniqueness, Bayes one-step Markov, hybrid multistep Markov, compression, IPAM and sequence-match to evaluate their effectiveness in anomaly detection. In [22], the authors propose Naive Bayes classifier on a data set containing truncated user commands for masquerade detection, it is based on the assumption that a user generates a command with a fixed probability which is independent of the command preceding it. The authors in [23] propose an updating method called adaptive Naive Bayes algorithm incorporating deferred decision, which use a progressively varying threshold for consecutive masquerade blocks. To improve the analysis efficiency, the authors in [24] employed non-negative matrix factorization (NMF) to reduce high dimensional data for intrusion detection with high efficiency and low use of system resources. The authors in [25] use the frequencies of system calls and commands to model the program or user behaviors, then they apply the principal component analysis (PCA) for dimensional reduction and anomaly detection. The above methods can extract effective features and perform abnormal behavior detection, but it is not enough to just use the transition attributes or the frequency attributes to model the behaviors, how to fully utilize the data and better characterize user behaviors to help the administrators perform system security management is still to be explored.

Enlightened by the related work, we propose the BR model to characterize and visualize the user's behavior from massive system logs. Based on the BRs, we conduct the clustering and NMF analysis to mine the user clusters and anomalies to improve the system security management efficiency.

III. THE FRAMEWORK AND MODELING

A. FRAMEWORK OF THE PROPOSED METHODS

The framework for user behavior visualization and system security management is divided into two parts. The first part focuses on user behavior visualization and management, the second part is for system security management, which focuses on abnormal behavior detection and fast anomaly tracing, as shown in Figure 1.

Step 1: Log collection. Multi-type logs are collected using Linux syslog mechanism from the remote servers which provide different kind of services [26]. The logs collected include the message and secure log from the servers running in actual network environment and the labeled anomaly logs from the target server.

Step 2: BR model establishment for behavior visualization. Based on the logs collected, we construct the BR model by characterizing the distribution of the commands during different time windows, which can reflect the user's behavior characteristics in a visual way.

Step 3: BR clustering for user behavior management. We employ the matrix to quantify the BR and apply DBSCAN clustering method to the BR matrices, in this way, we can classify the users with similar behavior patterns into the same cluster and design different management policies to different clusters and improve behavior management efficiency.

Step 4: Abnormal behavior detection and tracing for system security management. We apply the NMF method to the BR matrices and perform abnormal behavior detection, and then a fast anomaly tracking method for potential threat mining is proposed. By mining and controlling the abnormal behaviors, we can keep the system under control and improve the service stability.

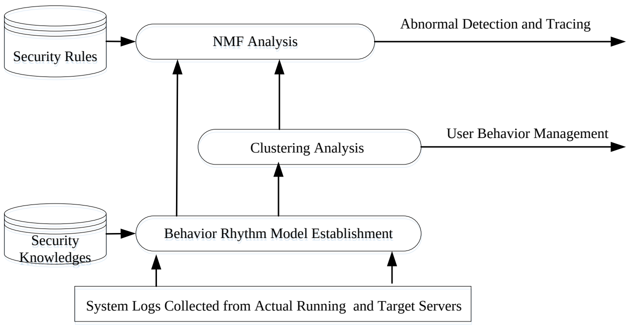


Figure 1. Framework of the proposed methods

B. PLATFORM FOR MASSIVE LOG COLLECTION

To bridge the gap between research in academia and practice in industry, we construct a log collection and system security platform which integrates with parallel computing technologies and big data storage mechanism to process massive logs effectively [27-28], the structure of the platform is shown in Figure 2.

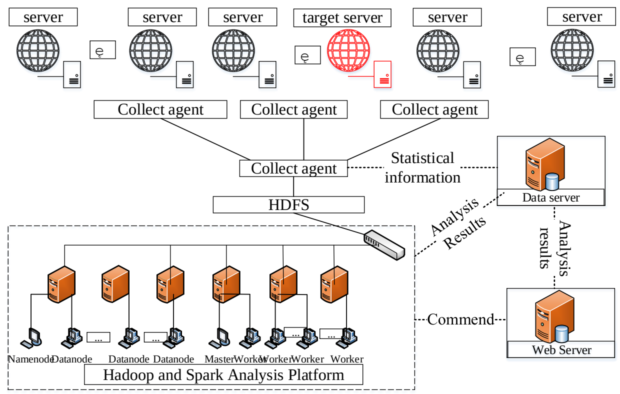


Figure 2. The structure of security management platform

The platform collects system logs from more than 50 servers and 1 target server located in the campus network of Xi'an Jiaotong University. The servers are selected from the campus network center, which are mainly used by colleges or institutions to establish their own websites, providing news and email services for their employees. The target server, which is marked as red in Figure 2, has same hardware and software configurations with the actual servers, and engineer attack the target server to generate labeled logs. The Web management server is responsible for human-computer interaction, which can receive the analysis instructions from the administrators and send the commands to the platform for appropriate operations, and then visualize the final detection results. We employ Flume to collect massive data and store them into HDFS (Hadoop Distributed File System), and introduce Spark technology into the platform for real-time massive data processing, which also provides the platform with the scalability and fault tolerance capabilities.

C. BEHAVIOR FEATURES EXTRACTION

System logs record all kinds of user behaviors and are one of the important data sources for user behavior monitoring and system security management. In this paper, we mainly use the secure and the message log to profile user's behaviors. The secure log records the user activity information, including the logging time, logging IP address, user name, login successful or not and so on. The message log records the related information about the commands used in the system, including time, IP and command series. We give the log examples which are shown in Table 1.

Table 1. Different types of log examples

Log type	Log samples
Secure	Jan 7 15:32:09 linux logserver sshd[2668]: Failed password for user2 from 10.220.1.20 port 52048 ssh2 Jan 7 16:27:38 linux logserver sshd[3066]: Accepted keyboard interactive/pam for user2 from 10.220.1.20 port 52121 ssh2 Jan 7 15:32:37 linux-logserver sshd[1153]: Received disconnect from 10.220.1.20: 11: disconnected by user
Message	Jan 7 13:07:20 linux-logserver user2_shell_cmd: [/home/user2] 360 [2016-01-07 13:07:20] [10.220.1.20] cd /var/log/ Jan 7 13:07:23 linux-logserver user2_shell_cmd: [/var/log] 361 [2016-01-07 13:07:23] [10.220.1.20] ls

To fully explore the users' behavior characteristics, we need extract features which can reflect the users behavior characteristics, based on the management experiences, we mainly extract three features listed below:

Feature 1: The IP address. The IP address is used to identify which host the user used to connect the server and this feature can be extracted from the secure log. We can obtain the corresponding physical location of the user based on the IP address, and use it for preliminary abnormal detection and access control. In this paper, we mainly divide the IP addresses into three categories, including the IP addresses belong to the CERNET, the other IP addresses in China and the foreign IP addresses. The network administrators usually use the Dynamic Host Configuration Protocol (DHCP) technology to making effective use of Internet address, and use Network Address Translation (NAT) technology to translate the private IP addresses to the public IP addresses, which works fine for our method since the public IP addresses can also help us to get the IP category information.

Feature 2: The time window when the activity happened. This feature can be extracted from the secure log. In this paper, we divide one day into 24-time windows, each time window represents one hour. We find that users usually have routine behavior patterns, they use similar commands during the same time window of different days due to their behavior habits, thus this feature can be used to characterize the user's behavior.

Feature 3: The command type used. We can extract the commands from the message log, the command series are closely tied to user behaviors, and from the command type used we can infer the user's operation purpose. Generally speaking, there are mainly two kinds of users in one system, including the users responsible for system security maintenance and the users who use the services provided by the system, they use different commands for different purposes. We divide the system commands into 8 categories based on their functions, the detailed descriptions are shown in Table 2.

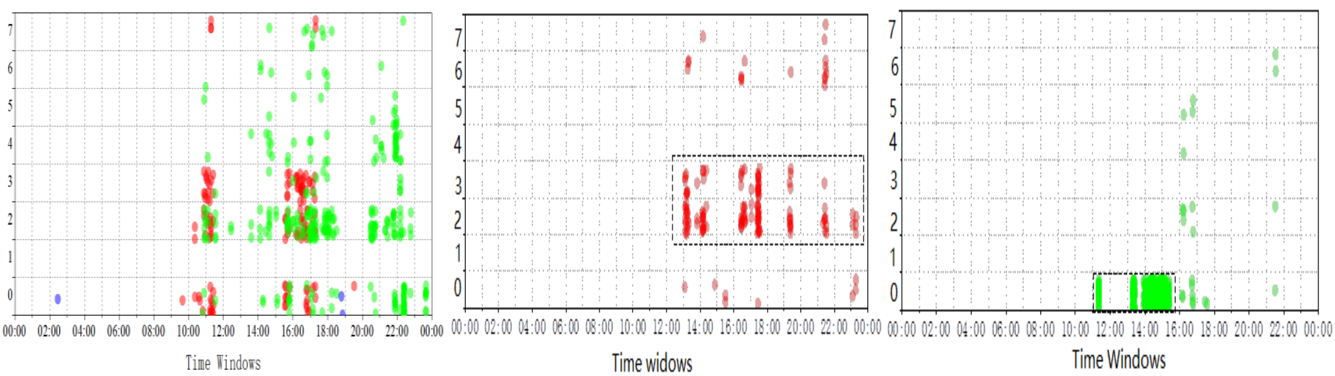
Table 2. Categories of the command

No.	Category	Command examples
0	Logging and installation	login, poweroff, shutdown
1	Graphical interface	setup, gdmsetup, gdmconfig
2	Information display	time, data, cal
3	File operations	file, mkdir, grep
4	System management	gnome-terminal, mkfs,
5	Network management	ifconfig, ping, route
6	Security operations	password, chown, chmod
7	Others	xine, xmms, bc, etc.

D. BEHAVIOR RHYTHM MODEL

To present the user's behavior characteristics in an easily understandable way, a new model named as Behavior Rhythm (BR) is proposed. We use red, green, and blue dots to characterize the behavior characteristics from different types of logging IP addresses, red means the user login to the system from the CERNET, while green represents that of other IP addresses in China, and blue represents that of foreign IP addresses. Each dot represents the user uses a specific type of commands at the specific time window. As shown in Figure 3, the X axis represents the time windows ranging from 0 to 24, and the Y axis represents different kinds of commands. The density of the dots represents the frequency of operations in the specific time window. To avoid overlap among different dots, we select the detailed time points when the activity occurred to locate the X-coordinate, while locating the Y-coordinate randomly in the corresponding small square. If there is no activity captured, the corresponding square is blank.

We select three typical BRs as an example to verify its efficiency in behavior profiling as shown in Figure 3. From the figures, we can obtain several advantages of this model. Firstly, the model can characterize the complex operations efficiently by giving a macroscopical behavior sketch, we can obtain the most frequent commands used and their corresponding time windows to infer the user's role. The model can also reflect the characteristics of the users from different types of logging IP addresses, e.g. Figure 3a is a system administrator who frequently uses information display and system configuration operations to keep the system under control, the user mainly uses the services provided by the CERNET and other IP addresses in China. Secondly, the model can reflect the users' behavior habits and their daily lives, e.g. Figure 3b is the BR of a software engineer who mainly works in the afternoon. Finally, the BR model can be used to mine some obvious anomalies, such as in Figure 3a, there are some logging attempts from the IP address outside China, which are denoted by blue dots in the BR, the logging attempts may be generated by hackers. In Figure 3c, there are lots of logging related commands performed in a short time window as marked by the dotted rectangle, which is an obvious sign of password crack attack.



a) BR of user using multi-IP types b) BR of engineer c) BR of password crack attack

Figure 3. Examples of typical BRs

IV. BR CLUSTERING AND ITS APPLICATION TO SYSTEM MANAGEMENT

A. BR Clustering using DBSCAN

Based on the BR definition, we can employ a matrix to quantify the BR, there are 8*24 small squares in the BR, we regard each square in the BR as an element in the matrix, but each element contains three frequency values due to the three types of IP addresses, thus the BR can be converted to a 3*8*24 matrix. Here we expand the BR matrix to 24*24 based on the types of IP address as follows, the first 8 rows are for CERNET, and the second and the last 8 rows are for the other IP addresses from China and the foreign IP addresses respectively, we employ the Equation 1 to denote it, where $v_{i,j}$ ($0 \le i \le 24$, $0 \le j \le 24$) denotes the frequency of the i-th command type from the [i/8]-th type of IP addresses during the j-th time window.

$$BR=\begin{bmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,24} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,24} \\ \cdots & \cdots & \cdots & \cdots \\ v_{24,1} & v_{24,2} & \cdots & v_{24,24} \end{bmatrix} \quad (1)$$

In order to perform clustering and find the users with similar behavior characteristics, we convert the BR matrix into a vector with 576 attribute values and then apply the clustering method to the vectors, in this way we can design different policies to different clusters to improve the network management efficiency. In this paper we choose DBSCAN [1] algorithm because it can identify arbitrary shaped clusters and eliminate noise data. The detailed pseudocode is shown as follows:

Input: BR sets $D=\{BR_1,...,BR_m\}$, parameters (ε, MinPts)

Output: Clusters $C=\{C_1,C_2,...,C_k\}$

Initialize the core object sets $\Omega=\phi$
For j = 1 to m:
    Find the ε-neighbors $N_\varepsilon(x_j)$ for x_j
    If $|N_\varepsilon(x_j)|\geq MinPts$:
        Insert x_j into the object set: $\Omega=\Omega\cup\{x_j\}$
Initialize number of clusters k = 0
Initialize sample sets $\Gamma=D$
While $\Omega\neq\phi$
    Mark the items which do not get processed $\Gamma_{old}=\Gamma$
    Randomly select $o\in\Omega$, initialize $Q=<o>$
    $\Gamma=\Gamma\setminus\{o\}$
    While $Q\neq\phi$
        Extract the first sample q from Q
        If $|N_\varepsilon(q)|\geq MinPts$
            Let $\Delta=N_\varepsilon(q)\cap\Gamma$
            Insert the samples in Δ into Q
            $\Gamma=\Gamma\setminus\Delta$
        $k=k+1$, Obtain the clusters $C_k=\Gamma_{old}\setminus\Gamma$
    $\Omega=\Omega\setminus C_k$

B. FEATURES FOR DEEP CLUSTER INVESTIGATION

To deeply investigate the behavior characteristics for the users in different clusters, we define the following two features to capture the behavior characteristics in different clusters, which can help the administrators to design different access control or management policies for different types of users to improve the system management efficiency.

1) OPERATION AND MAINTENANCE FREQUENCY

We employ the Operation and Maintenance Frequency (OMF) to measure the frequency of the operations of the specific type of users, which is defined in Equation 2. The OMF is not only related to the number of login per unit time, but also related to the operation frequency after user login.

$$OMF=\sum_{i=1}^{N}\frac{\log M_i}{T_i} \quad (2)$$

Where N is the number of logging times, M_i is the number of commands for i-th login, T_i is the session lifetime for i-th login. The frequency is proportional to the number of user's commands and inversely proportional to the length of session lifetime, and we take a logarithm of M_i to eliminate the effect of excessively large number of commands. If the user frequently login to the system and perform a large number of commands during each session, the OMF of the user will be high. On the contrary, if the user does not operate much during each session, the OMF of the user will be low.

2) FREQUENT COMMAND SEQUENCE

System commands are the most basic way for a user to interact with the system. Analyzing the users' frequent command sequences is helpful to discover users' behavior patterns, which can be used to reflect the user's customary behaviors or to infer the user's role. We can get the command sequences that user performed from the message log, and then Prefixspan algorithm [2] is applied to mine the frequent command sequences from all the command sequences the users performed. If the frequent command sequence mined for the user belongs to the network related category, then the user is very likely to be a network engineer. Here we apply Prefixspan algorithm to all the command series of the users in the specific cluster to mine the frequent command sequences for the cluster, then use them to infer the users' role in the cluster.

V. RHYTHM DECOMPOSITION AND ITS APPLICATION TO ABNORMAL DETECTION

A. ABNORMAL DETECTION USING NMF

From the above discussion, we can see that different users have different BRs, thus the BR of the anomalies should be different with the BRs of the normal users. We apply NMF method to decompose the BR and then perform abnormal detection. NMF is a powerful method for reducing representation of data and has been successfully applied to face recognition, document classification, and abnormal detection, etc [29]. For a given matrix X, where each column is an m-dimensional non-negative vector of the original data (n vectors). Based on the theory of NMF [3], we can find two new matrices W and H, to approximate the original matrix, which can be denoted in Equation 3:

$$\mathrm{X}_{m\times n}\approx\mathrm{W}_{m\times r}\mathrm{H}_{r\times n} \quad (3)$$

In this way, the data matrix X is factorized into two smaller matrices with positive elements, the matrix W as the base and the other matrix H as the encoding coefficients representing normal or abnormal features, and we employ them to perform abnormal detection.

To obtain the solution for Equation 3, an interactive method is given by the following rules [3]:

$$\mathrm{H}_{\mathrm{rn}} \leftarrow \mathrm{H}_{\mathrm{rn}} \frac{(\mathrm{W}^{\mathrm{T}}\mathrm{X})_{\mathrm{rn}}}{(\mathrm{W}^{\mathrm{T}}\mathrm{WH})_{rn}} \tag{4}$$

$$W_{\mathrm{mr}} \leftarrow W_{\mathrm{mr}} \frac{(XH^T)_{mr}}{(WHH^T)_{mr}} \tag{5}$$

The initialization of W and H is performed by random selecting positive initial conditions, and the convergence of the process is also ensured. Based on the theory of NMF, we can find the matrices W and H to denote the BR, as shown in Equation 6.

$$BR_{24\times 24} \approx W_{24\times r} H_{r\times 24} \tag{6}$$

Based on the property of frequency, the following normalization condition is satisfied:

$$\sum_{i=1}^{24} v_{i,j} = 1 \qquad (1 \le j \le 24) \tag{7}$$

Based on the iterative algorithm for solving the NMF problem, we can get the following property:

$$\sum_{i=1}^{r} h_{i,j} = \sum_{i=1}^{24} v_{i,j} = 1 \qquad (1 \le j \le 24) \tag{8}$$

Which shows that the sum of the elements in each column of H is equal to 1, in other words, any normal data can be characterized by a single value 1 by adding all the elements in each column vector h of H. If the specific column vector h corresponds to normal behaviors, its associated coefficient vector should have similar characteristics. In other words, if the data vector t corresponds to normal behaviors, the sum of all the elements in h_t should be approximately equal to 1. Thus we define the following simple classifier as the anomaly degree for abnormal detection. If ε_j is above the threshold δ, the user behavior associated with the specific column vector h_j is considered as abnormal behavior, otherwise it is treated as normal behavior.

$$\left| \sum_{i=1}^{r} h_{i,j} - 1 \right| = \varepsilon_j \qquad (1 \le j \le 24) \tag{9}$$

As the BRs are constructed to reflect the user's complete behavior characteristics for each day, some columns have no values as there are no operations during that time period, we need to mark them before performing decomposition. The detailed process of anomaly detection using NMF can be described as follows:

Step 1: We mark all the columns in the BRs whose sum equal to zero, we sum all the columns before performing analysis and get $Z = (z_1, z_2 .... z_{24})$, where $z_j = \sum_{i=1}^{r} v_{i,j} \ (1 \le j \le 24)$.

Step 2: We apply NMF method to decompose the BR matrix into two smaller matrices W and H, and we denote the coefficients matrix H using $H = (h_{r,1}, h_{r,2} .... h_{r,24})$.

Step 3: Calculate the sum of each column of the coefficients matrix H and get $S = (s_1, s_2 .... s_{24})$, where $s_j = \sum_{i=1}^{r} h_{i,j} \ (1 \le j \le 24)$.

Step 4: The sum of each column of the coefficients matrix H is employed to denote the abnormal degree of the BR during the specific time window. To eliminate the influence of the columns without operations, the columns of the original BR matrix whose sum equal to zero need to be excluded, then we calculate the abnormal degree $D = (d_1, d_2 .... d_{24})$ as following:

$$d_j = \begin{cases} s_j, \ s.t.\ z_j \ne 0 \\ 1, \quad s.t.\ z_j = 0 \end{cases} \quad (1 \le j \le 24) \tag{10}$$

Step 5: we employ the value of $\varepsilon_j = |1\text{-}d_j| \ (1 \le j \le 24)$ to judge whether the behaviors during the j-th time window of the BR are abnormal or not, if ε_j is bigger than the threshold δ, we claim the behaviors are abnormal.

B. FAST ANOMALY TRACKING

Fast anomaly tracking is very important for keeping system normal running, we can control the anomalies in time and reduce their influence on the system. In this paper, the BR model reflects the user behaviors by describing the frequency of the commands used based on the dimensions of time and function category. Based on this property, we propose a fast tracking method based on the BR similarity calculation, we first convert the BR into a vector with 576 attribute values, and then use the Cosine Distance to measure the similarity between the template BR and the BRs in the actual network as shown in Equation 11. If the similarity is very high, the BRs can be marked as anomalies corresponding to the template BR. In this way, the administrators can use the BRs of typical anomalies as templates, or even design the template BRs based on the abnormal behavior characteristics and their management experiences, to track the potential threats they are interested in efficiently from the actual network environment.

$$dist(X,Y)=1-\frac{\sum_{i=1}^{576} x_i y_i}{\sqrt{\sum_{i=1}^{576} x_i^2}\sqrt{\sum_{i=1}^{576} y_i^2}} \tag{11}$$

VI. EXPERIMENTAL RESULTS

A. DATA COLLECTION

Based on the platform constructed in the campus network center, we mainly collect two kinds of system logs. One is the system logs from the actual servers and the other is the labeled anomaly logs from the target server. The target server is used to generate labeled anomaly logs, which has the same configuration with the other actual servers, only the access control strategy of the target server is less restricted compared to the other actual servers. All the logs are collected from these servers during the period from November 2015 to April 2016. During data collection phase, the target server has been attacked by the engineers from our Lab, Huawei's security engineers and Dutch hackers. Some typical intrusion behaviors such as illegal log modification, illegal authority promotion, are performed on the target server to construct the benchmark data. According to the relative information of the attack events, such as the attack time and account, the logs corresponding to the specific abnormal behaviors can be easily marked. The basic information of the system logs collected is shown in Table 3 and Table 4.

Table 3. The logs collected from actual server

Log type	# of Records	Servers	Time
Message	2430479721	219.245.37.1~219.245.37.68	2015.11-2016.04
Secure	69372234	219.245.37.1~219.245.37.68	2015.11-2016.04
Total	2499851955	219.245.37.1~219.245.37.68	2015.11-2016.04

Table 4. The logs collected from target server

Log type	# of Records	Servers	Time
Message	82462376	202.117.54.250	2015.11-2016.04
Secure	97318864	202.117.54.250	2015.11-2016.04
Total	381434885	202.117.54.250	2015.11-2016.04

Based on the system logs collected, we extract the corresponding features and construct the BRs for users. The BRs obtained from the target server are labeled as anomalies, include the BRs of typical attacks, such as the password crack attack, authority promotion, etc. We totally obtained 2111 BRs, including 101 labeled abnormal BRs, we divided the whole dataset into three sub-datasets to provide different proportions of abnormal versus normal BRs, and then employ these sub-datasets to evaluate the performance of our methods in abnormal detection. The statistic information of the sub-datasets is shown in Table 5.

Table 5. The statistics of datasets

No.	# of Total BRs	# of Normal BRs	# of Abnormal BRs
Dataset 1	704	670	34
Dataset 2	737	670	67
Dataset 3	771	670	101

B. ANALYSIS ON THE CLUSTERING RESULTS

We apply DBSCAN algorithm to the sub-datasets and obtain average 13 clusters in different datasets at different time points, as shown in Figure 4a, we can find that the number of clusters is very stable along with the time changing. The users in each cluster have similar behavior characteristics, they use the same types of commands during the similar time window to perform similar tasks. Figure 4b analyzes the stability of the clustering results, which shows about 85% of the total users in the datasets keep staying in the same cluster over time, which indicates the high stability of BR clusters during a long time period. These observations confirm that the BR can capture the user's behavior characteristics and separate users into the stable clusters. We apply the DBSCAN methods to the whole data set and then analyze the number of users in different clusters, and the analysis results are shown in Figure 4c, we can find that the top 3 big clusters occupy most of the users, according to the network management experiences, most of the time the servers are running in the normal state, so most of the BRs are generated by normal behaviors, so we can regard those big clusters as normal users who have similar roles or purposes. There are also some isolated nodes containing only one user, which may be caused by abnormal behaviors. In this paper, we choose ε=0.65 and MinPts=10 for the clustering analysis.

With the clustering results, we can deeply explore the cluster characteristics by analyzing the OMF and the frequent command sequence for the users in the specific cluster, and then design suitable management policies for different kinds of users to improve the management efficiency. We mainly analyze the top 3 clusters which occupy most of the users and one small cluster as example, and select the cluster center as the representative BR of the cluster, the results are shown in Figure 5.

For the biggest cluster, we find that the OMFs of the users are not high, the frequent commands are related to information display and file operation, but do not involve the system management and network operations, and the users only use one type of logging IP address type which is the CERNET. We obtain two frequent command sequences [ls -al] and [vi], which are used to check the files and edit the code. Thus we infer this kind of users should be the students, who mainly stay in the LAB and use the CERNET network to do some algorithm tests during the daytime, Figure 5a is the representative BR of this cluster.

For the second biggest cluster, the OMFs of the users are higher than the first cluster, besides some information display and file operations, the frequent commands mainly involve the system management and network operations, the users only logged in from the CERNET. We obtain three frequent command sequences [netstat -na], [nc] and [ss -ant], which are used to check the current socket statistics and basic network-related information. Thus we infer that this kind of users should be system maintenance engineers, who need to configure the network and maintain the systems to provide better services to satisfy the needs of users, Figure 5b is the representative BR of this cluster.

It then follows another cluster, in which the users logged from multiple types of IP addresses, the users perform different operations using different types of IP addresses, as shown in Figure 5c, which is the BR of a cooperation engineer from Huawei company, who logins to the system to perform some configuration operations from different network environments. Most of the time he operates the system from Huawei company, but sometimes he comes to Xi'an Jiaotong University for code debugging, which can be seen from the two dotted-line rectangles. We also analyzed one small cluster and the representative BR is shown in Figure 5d. We find its OMF is very low and the operations are extremely simple, it is likely that these users are testing accounts, which are created to test whether the unauthorized operations or attacks are successful or not.

From the above analysis we can find different clusters have different behavior characteristics, we can design different access control policies for efficient security management, e.g. for the testing accounts we just mentioned, their authorities can be strictly limited like they are not allowed to perform the system management and network operations.

C. ANALYSIS ON THE NMF DECOMPOSITION RESULTS

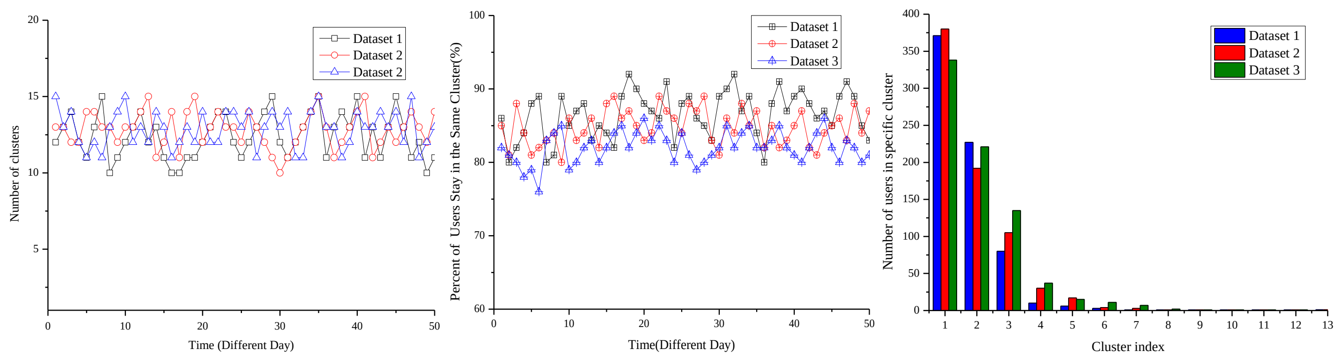
1) PERFORMANCE EVALUATION ON THE ABNORMAL DETECTION

We use the detection rate and false alarm rate, which are widely used in the related literatures [24], to evaluate the performance of the proposed method. Their definitions are given in Equation 11 and 12.

$$Detection\ Rate\ (DR) = \frac{TP}{TP + FN} \tag{12}$$

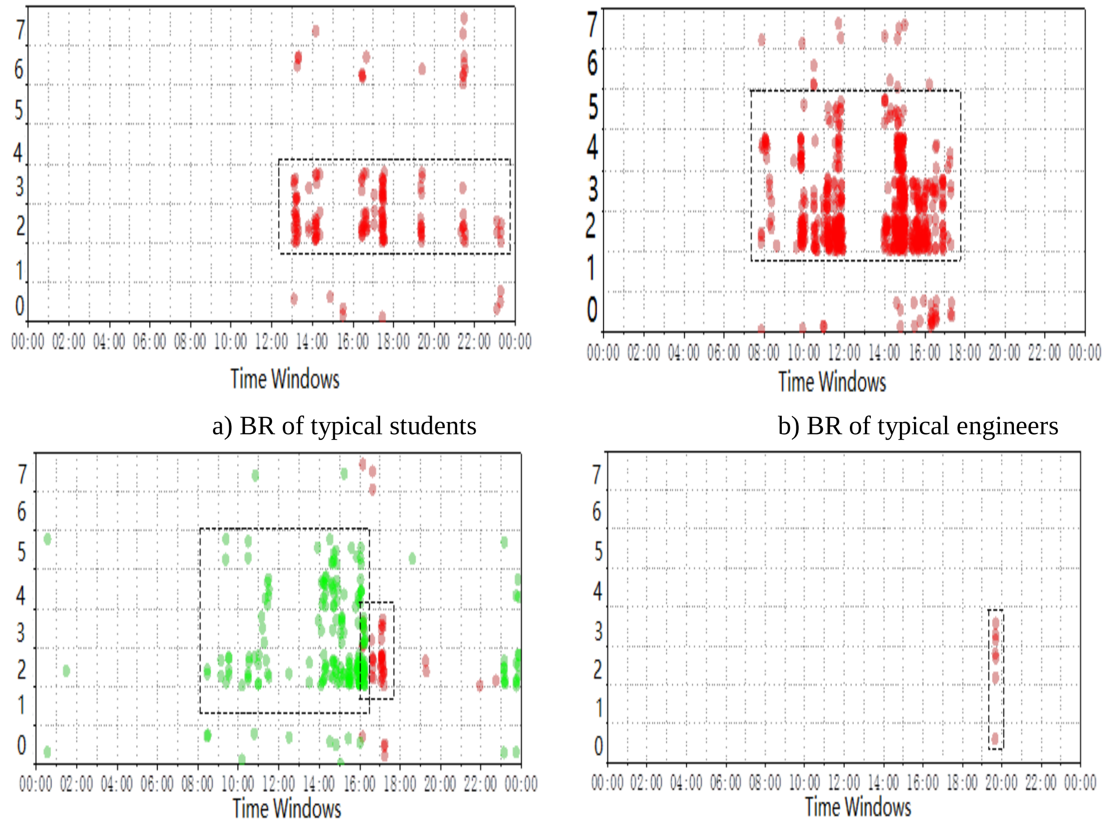
$$False\ Alarm\ Rate\ (FAR) = \frac{FP}{FP + TN} \tag{13}$$

where True positives (TP) denotes the number of the anomalies which correctly classified as anomalies, True negatives (TN) denotes the number of the normal events which correctly classified as normal events, False positives (FP) denotes the number of the normal events which wrongly classified as anomalies and False negative (FN) represents the number of the anomalies which wrongly classified as normal events.



a. Number of clusters　　b. Stability of clustering results　　c. User distribution with clusters

Figure 4. The analysis of the clustering results



a) BR of typical students
b) BR of typical engineers
c) BR of cooperation engineer
d) BR of testing accounts

Figure 5. BRs from different clusters

We employ two different ways to evaluate our method. Firstly, we select the main idea proposed in [25] to construct the comparison method and perform evaluation use the same data set with our method (denoted by NMF-BR). In [25], the authors employ the Principal Component Analysis (PCA) for dimension reduction and then employ the principal components extracted to train a model and perform abnormal detection. In this paper, we employ the PCA method to analyze the BRs obtained and then perform anomaly detection (denoted by PCA-BR). We select r=8 in Equation 3 and δ=0.3, 0.35, 0.4 as the thresholds for the proposed method, the evaluation results are summarized in Table 6. We can find that the proposed method shows better performance than the PCA-BR method in terms of both detection rates and false alarm rates.

Table 6. Evaluation results with the BR datasets

Methods	Dataset 1		Dataset 2		Dataset 3	
Metrics	DR	FAR	DR	FAR	DR	FAR
PCA-BR	76.47%	2.24%	82.09%	2.84%	80.20%	2.54%
δ=0.3	88.24%	4.18%	89.55%	3.58%	90.10%	3.73%
δ=0.35	85.29%	3.28%	88.06%	3.43%	89.11%	3.58%
δ=0.4	85.29 %	1.79 %	83.58%	1.94%	85.15%	2.09%

Secondly, we compare the proposed method with the method proposed in [24] and the evaluation results are shown in Table 7. The authors in [24] mainly use the frequency attributes of command data as a static model, then apply NMF algorithm for anomaly intrusion detection, we use NMF-FA to denote it. Here we choose r=8 in Equation 3 and δ=0.4 as the threshold for the proposed method to perform abnormal detection. From Table 7, it is seen that our method has better performance than the method proposed in [24]. The results verify the BR model can better capture user behavior characteristics, we not only consider the frequency attributes, but also combine the time dimension and the logging IP address information extracted from the related logs. In addition, our method does not need training process and prior knowledge, which greatly improve its usability.

Table 7. Evaluation results with different datasets

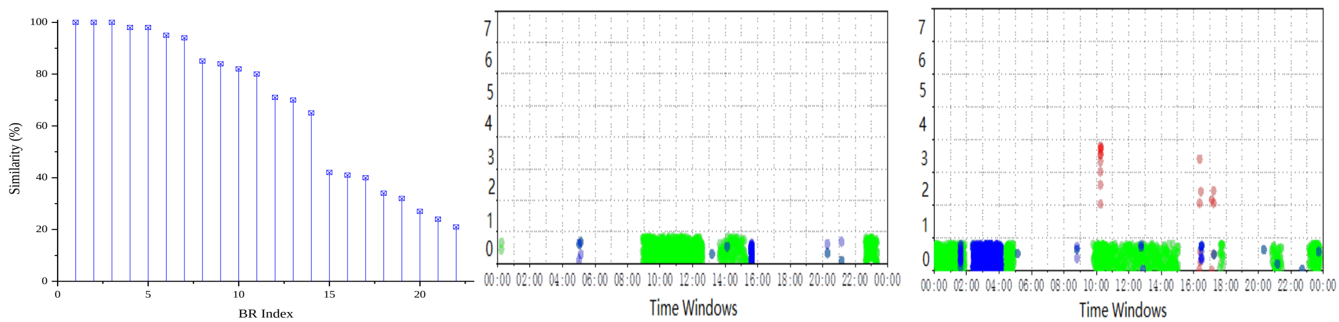
Methods	Dataset 1		Dataset 2		Dataset 3	
Metrics	DR	FAR	DR	FAR	DR	FAR
NMF-FA	67.65%	2.39%	70.15%	2.24%	71.29%	2.54%
NMF-BR	85.29 %	1.79 %	83.58%	1.94%	85.15%	2.09%

2) FAST ANOMALY TRACKING

Tracing the interested BRs is an important way for the administrators to mine the potential threats, we give an example to explain the fast anomaly tracking method we proposed. Here we try to find the users suffered from password crack attacks, so we choose a labeled BR as template, then we employ the cosine distance to calculate the similarity between the template BR and the BRs obtained from the actual network. The analysis results are shown in Figure 6. Figure 6a is the similarity results with the 22 most similar BRs sorted in descending order, and the top two similar BRs are shown in Figure 6b and Figure 6c. As Figure 6a shows, when we set the similarity threshold as 0.80, we can get 11 BRs, in which there are 10 labeled abnormal BRs, achieve 90.9% precision. In this way the administrators can find the similar threats from the actual network efficiently, and then they can give these accounts security limits to reduce the risks posed by these attacks timely.

VII. CONCLUSION AND FUTURE WORKS

In this paper, we present a framework for user behavior visualization and efficient system security management. Firstly, we propose the Behavior Rhythm model to reflect the user's behavior in a visual and easily understandable way. By analyzing the characteristics of the Behavior Rhythm, we can obtain the user's behavior habits, e.g. the type of commands they usually performed and the time windows they tend to login. We apply DBSCAN clustering method to the BRs and separate them into different clusters. The users in the same cluster have similar characteristics, and more than 85% of the users stay in the same cluster with time changing, which verify the stability of the users' behaviors, thus we can design different manage policies to the users in different clusters to improve the management efficiency. We apply NMF method to the BRs and perform abnormal detection based on the decomposition results. Based on the characteristics of the abnormal BRs, we provide a fast anomaly tracking method based on the similarity calculation, in this way the administrator can mine the potential threats efficiently and control the anomalies at the early stage, to improve the system stability and quality of service. For future works, we will focus on the fine-grained command classification and clustering results measurement.



a) BR similarity b) The BR traced c) The BR traced

Figure 6. Fast anomaly tracking

REFERENCES

[1] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with Noise," in *Proc. International Conference on Knowledge Discovery and Data Mining*, Portland, OR, 1996, vol. 96, no. 34, pp. 226-231.

[2] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu, "Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth," in *Proc. International Conference on Data Engineering*, Heidelberg, Germany, 2001, pp. 215-224.

[3] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, 2001, pp. 556-562.

[4] Apache Flume. Accessed: Aug. 2016. [Online]. Available: https://flume.apache.org/

[5] Apache Hadoop. Accessed: Aug. 2016. [Online]. Available: http://hadoop.apache.org/

[6] Apache Spark. Accessed: Aug. 2016. [Online]. Available: https://spark.apache.org/

[7] R. Vaarandi, and M. Pihelgas, "Using security logs for collecting and reporting technical security metrics," in *Proc. IEEE Military Communications Conference*, Baltimore, MD, USA, 2014, pp. 294-299.

[8] M. G. Schultz, E. Eskin, E. Zadok, and S. J. Stolfo, "Data mining methods for detection of new malicious executables," in *Proc. IEEE Symposium on Security and Privacy*, Oakland, CA, USA, 2001, pp. 38-49.

[9] N. Busany and S. Maoz, "Behavioral log analysis with statistical guarantees," in *Proc. International Conference on Software Engineering*, Austin, TX, USA, 2016, pp. 877-887.

[10] J.-G. Lou, Q. Fu, S. Yang, Y. Xu, and J. Li, "Mining invariants from console logs for system problem detection," in *Proc. USENIX Annual Technical Conference*, 2010, pp. 231-244.

[11] T. Kimura, K. Ishibashi, T. Mori, H. Sawada, T. Toyono, K. Nishimatsu, and A. Watanabe, "Spatio-temporal factorization of log data for understanding network events," In *Proc. IEEE Conference on Computer Communications*, Toronto, ON, Canada, 2014, pp. 610-618.

[12] J. Ya, T. Liu, Q. Li, J. Shi, H. Zhang, P. Lv, and L. Guo, "Mining host behavior patterns from massive network and security logs," in *Procedia Computer Science*, 2017, vol. 108, pp. 38-47.

[13] S. Mehnaz and E. Bertino, "Building robust temporal user rhythms for anomaly detection in file system accesses," in *Proc. Annual Conference on Privacy, Security and Trust*, Auckland, New Zealand, 2017, pp. 207-210

[14] W. Xu, L. Huang, A. Fox, D. Patterson, and M. Jordan, "Mining console logs for large-scale system problem detection," in *Workshop*

on Tackling Computer Systems Problems with Machine Learning Techniques, 2008, pp. 4-4.

[15] W. Xu, L. Huang, A. Fox, D. Patterson, and M. I. Jordan, "Detecting large-scale system problems by mining console logs," in *Proc. ACM SIGOPS Symposium on Operating Systems Principles*, 2009, pp. 117-132.

[16] Q. Fu, J.-G. Lou, Y. Wang, J. Li, "Execution anomaly detection in distributed systems through unstructured log analysis," in *Proc. IEEE Conference on Data Mining*, Miami, FL, USA, 2009, pp. 149-158.

[17] K. Xu, Z. Zhang, S. Bhattacharyya, "Internet traffic behavior profiling for network security monitoring, " in *IEEE/ACM Transactions on Networking (TON)*, 2008, vol. 16, no. 6, pp. 1241-1252.

[18] K. Xu, F. Wang, L. Gu, "Behavior analysis of Internet traffic via bipartite graphs and one-mode projections," in *IEEE/ACM Transactions on Networking (TON)*, 2014, vol. 22, no. 3, pp. 931-942.

[19] D. Zhao, I. Traore, B. Sayed, W. Lu, S. Saad, A. Ghorbani, D. Garant, "Botnet detection based on traffic behavior analysis and flow intervals," in *Computers & Security*, 2013, vol. 39, pp. 2-16.

[20] T. Qin, X. Guan, C. Wang, Z. Liu, "MUCM: multilevel user cluster mining based on behavior profiles for network monitoring," in *IEEE Systems Journal*, 2015, vol. 9, no. 4, pp. 1322-1333.

[21] M. Schonlau, W. Dumouchel, W.-H. Ju, A. F. Karr, M. Theus, and Y. Vardi, "Computer intrusion: Detecting masquerades," in *Statistical Science*, 2001, vol. 16, no. 1, pp. 58-74.

[22] R. A. Maxion and T. N. Townsend, "Masquerade detection using truncated command lines," in *Proc. International Conference on Dependable Systems and Networks*, Washington, DC, USA, 2002, pp. 219-228.

[23] S. K. Dash, K. S.Reddy, and A. K. Pujari, "Adaptive Naive Bayes method for masquerade detection," in *Security and Communication Networks*, 2011, vol. 4, no. 4, pp. 410-417.

[24] X. Guan, W. Wang, and X. Zhang, "Fast intrusion detection based on a non-negative matrix factorization model," in *Journal of Network and Computer Applications*, 2009, vol. 32, no. 1, pp. 31-44.

[25] W. Wang, X. Guan, X. Zhang, "Processing of massive audit data streams for real-time anomaly intrusion detection," in *Computer communications*, 2008, vol. 31, no 1, pp. 58-72.

[26] C. Lonvick, *The BSD syslog protocol*, document RFC 3164, 2001.

[27] T. Qin, Y. Gao, L. Wei, Z. Liu, and C. Wang, "Potential threats mining methods based on correlation analysis of multi-type logs," in *IET Networks*, 2017.

[28] Z. Liu, T. Qin, X. Guan, H. Jiang, and C. Wang, "An integrated method for anomaly detection from massive system logs," in *IEEE Access*, 2018.

[29] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," in *Nature*, 1999, vol. 401, 788-791.

ZHAOLI LIU received the B.S. degree and M.S. degree in computer science and technology from Xi'an Jiaotong University, Xi'an, China, in 2007 and 2010 respectively. She is currently pursuing the Ph.D. degree in computer science and technology at Xi'an Jiaotong University, and she is now a visiting PhD student in Prof. Weibo Gong's group at University of Massachusetts, Amherst, USA. Her research interests lie primarily in network security and online social network analysis.

XIAOHONG GUAN received the B.S. and M.S. degrees in control engineering from Tsinghua University, Beijing, China, in 1982 and 1985 respectively, and the Ph.D. degree in electrical and systems engineering from University of Connecticut, Storrs, USA, in 1993. He was a visiting professor with the Division of Engineering and Applied Science at Harvard University, Cambridge, USA, in 1999. He became an IEEE Fellow in 2007. He is currently a Cheung Kong Professor with the Department of Automation and the Dean of School of Electronic and Information Engineering. He was elected as an academician of the Chinese Academy of Sciences in 2017. His research focuses on scheduling of complex networked resources and network security.

SHANCANG LI received the B.S. and M.S. degrees in mechanics engineering and the Ph.D. degree in computer science from Xi'an Jiaotong University, Xi'an, China, in 2001, 2004, and 2008, respectively. He is currently a Senior Lecturer with the Department of Computer Science and Creative Technologies, University of the West of England, Bristol, U.K. His current research interests include digital forensics for emerging technologies, network security, cybercrimes, network attacks, wireless sensor networks and Internet of Things.

TAO QIN received the B.S. degree in information engineering and the Ph.D. degree in computer science and technology from Xi'an Jiaotong University, Xi'an, China, in 2004 and 2010 respectively. He is currently an Associate Professor with the Department of Computer Science and Technology and MOE KLINNS Lab, Xi'an Jiaotong University. His research focuses on anomaly detection and network traffic modeling.

CHAO HE received his B.S. degrees in computer science from Xi'an Jiaotong University, Xi'an, China, in 2015. He is currently a graduate student with the Department of Computer Science and MoE KLINNS Laboratory of Xian Jiaotong University. His research interests include data mining, user behavior analysis and abnormal detection.