IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 27, 2019, accepted April 20, 2019, date of publication May 1, 2019, date of current version May 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2914303

An Adaptive Profile-Based Approach for Detecting Anomalous Traffic in Backbone

XIAO-DONG ZANG, JIAN GONG, AND XIAO-YAN HU
[1]School of Cyber Science and Engineering, Southeast University, Nanjing 211189, China
[2]Jiangsu Provincial Key Laboratory of Computer Network Technology, Southeast University, Nanjing 211189, China
[3]Key Laboratory of Computer Network and Information Integration, Ministry of Education, Southeast University, Nanjing 211189, China

Corresponding author: Xiao-Dong Zang (xdzang@njnet.edu.cn)

This work was supported in part by the Jiangsu Key Laboratory of Computer Networking Technology and the Key Laboratory of Computer Network and Information Integration, Ministry of Education, Southeast University, in part by the National Natural Science Foundation of China under Grant 61602114, in part by the Key Research and Development Program of China under Grant 2017YFB0801703, and in part by the CERNET Innovation Project under Grant NGII20170406.

ABSTRACT Anomaly detection is the first step with a challenging task of securing a communication network, as the anomalies may indicate suspicious behaviors, attacks, network malfunctions, or failures. In this paper, we address the problem of not only detecting different anomalies, such as volume based (e.g., DDoS or Flash crowd) and spatial based (e.g., network scan), that arise simultaneously in the wild but also of attributing the anomalous point to a single-anomaly event causing it. Besides, we also tackle the problem of low-detection accuracy caused by the phenomenon of traffic drift. To this end, a novel adaptive profile-based anomaly detection scheme is proposed. More specifically, a more comprehensive metrics set is defined from the dimensions of temporal, spatial, category, and intensity to compose IP traffic behavior characteristic spectrum for fine-grained traffic characterization. Then, the digital signature matrix obtained by using the ant colony optimization (ACO) algorithm is applied to construct the baseline profile of the normal traffic behavior. Anomalous points are identified and analyzed by using confidence bands and a generic clustering technique, respectively. Finally, a lightweight updating strategy is applied to reduce the number of false positives. Real-world data of China Education Research Network backbone and synthetic data are collected to verify our proposal. The experimental results demonstrate that our approach provides a fine-grained behavior description ability and has significantly increased the detection accuracy compared with other similar alternatives.

INDEX TERMS Traffic characterization, anomaly detection, netflow, characteristic spectrum, digital signatures.

I. INTRODUCTION

Traffic anomalies or network anomalies [1], [2] refer to intentional or unintentional events in traffic flows that deviate from what is considered as normal in the context of network management. These unexpected events are generated by the damage ranging from network performance problems to security violations, including the results of spurious traffic caused by network failures, or suspicious behaviors such as network scans for vulnerable ports/services, attacks such as TCP SYN flooding and DDoS amplification attacks.

Traffic anomalies have become a troubling issue for both administrators and end users as they typically change the normal behavior in a malicious or unintentional manner, resulting in the congestion or the interruption of the availability of services on a network [3]. As a result, substantial losses may be caused to human or the organizations. Therefore, it is necessary to constantly monitor the traffic behavior of a network. Currently, there are two main methods for anomaly detection, signature-based and profile-based [4]–[7]. In the signature-based approach, it explores the prior knowledge of each kind of anomaly, and uses them to construct a feature database. Although, this approach could recognize the traffic anomaly with high accuracy, the prerequisite is that anomalous features are needed in advance, which hampers the ability of recognizing new anomalies. Therefore, profile-based anomaly detection approach is proposed, which constructs normal traffic behavior profile of a network segment by using

The associate editor coordinating the review of this manuscript and approving it for publication was Songwen Pei.

history traffic and treats any activity deviating from it as a possible intrusion [8].

As different anomalies like volume-based and spatial-based originated from various sources may arise at the same time in a real network scenario, detecting them by applying profile-based solutions with a single-dimension metrics [9]–[16] is not ideal or efficient. For example, [9]–[12] chose volume-based attributes, such as the number of bits, packets and flows, while, [13]–[16] used spatial-based attributes, including the source and destination IP addresses and ports. As for the former, although, they can match the basic behavior of DDoS attacks, flash crowds, they have the problem of identifying the anomalies that produce a small perturbation traffic (e.g. network scan). As for the latter, they can identify the anomalies like network scan with the spatial distribution features, however, they do not have the ability of identifying volume-based anomalies. In a word, the profile-based solutions with single-dimension metrics are inefficient and with lots of false negatives in the wild. What is more, the static nature of many preconceived baseline profiles prevent true adaptively due to the fact that real traffic evolves over time, namely, ''traffic drift''. Therefore, the baseline profile is needed to update without much computation, otherwise, false positives will be high if nothing is done when anomalies are slowly introduced in the profile and become legitimate events. To summarize, the main goal of this work is to answer the following questions:(1) Can we reliably detect the anomalies in a real network scenario that has anomalies responsible for generating a large number of traffic, along with others that produce a small perturbation traffic simultaneously? Also, can we attribute the detected anomalous points to a single anomaly event? (2) Can we reduce the number of false positives due to the result of traffic drift?

Aiming to address the questions mentioned above, in this paper, a novel anomaly detection system is proposed, which can identify the traffic anomalies and simultaneously capture other significant traits of the network. More specifically, a unidirectional IP flow record is considered as an independent IP activity. Defining the traffic behavior metrics of the IP activity from four dimensions, including the duration time, the peer addresses, the application types, and the number of packets and bytes contained in the flow, which correspond to the temporal, spatial, category and intensity, respectively. Nine single-attribute and thirty-nine dual-attribute metrics are extracted to compose an IP address traffic characteristic spectrum, so that the fine-grained traffic behavior of all IPs in the observed network can be captured. Then, ACO algorithm is used to obtain the digital signature matrix that is representative of the baseline profile of the network segment. Once the digital signatures of the normal traffic are captured, anomalous points are easy to detect by analyzing whether the disparities are higher or not than the predefined threshold. In order to minimize false positives, confidence bands are produced to restrict the interval where deviations are regarded as normal, and a proper correlation among these anomalous points is established to provide a better perspective of an event, so that unwanted notifications to administrator are reduced. Finally, a lightweight updating strategy is devised, hence, false positives rate can be reduced further. As a result, our work is not restricted to recognize the known and unknown anomalies, it can also provide fine-grained information of the network usage and users' behavior to the administrator. Compared with previous researches, the contributions of this paper include:

(1) A fine-grained and a more comprehensive traffic metrics set is proposed, which describes the traffic behavior of IP addresses from four dimensions including temporal, spatial, category and intensity.
(2) A concept of IP address traffic behavior characteristic spectrum is proposed that is the abstract of all possible characteristic values for each metric, and can be used to depict the behaviors of the IPs in the observed network.
(3) Providing a system that can detect different anomalies arising simultaneously in the wild autonomously, so that the administrator can be alerted when a possible problem is occurring.
(4) A lightweight updating strategy is applied to reduce false positives caused by the changes of the dynamic traffic, experimental results show that our work has the ability of identifying different anomalies arising at the same time in a real network scenario with substantially better detection accuracy and low computational complexity than other alternatives.

The remainder of the paper is organized as follows. In section 2, related works of anomaly detection approach are presented. In section 3, the details of the proposed anomaly detection approach are given. In section 4, we validate our approach by applying the described techniques on the backbone dataset and synthetic dataset. The conclusion of the paper is given in section 5.

II. RELATED WORKS

With the gradually expanding of network scale and the complexity of the network topological structure, monitoring the traffic behavior to ensure its availability and operability is becoming a tough task. Hence, a great number of anomaly detection approaches with many different techniques have been proposed in the past decade in order to meet the ever-increasing demands. We briefly describe the techniques of anomaly detection, and the significance of our work.

Network anomaly detection is an area of research being studied since 1980 [17], [18]. Currently, there are two main methods for anomaly detection, signature-based and profile-based. In signature-based approach, by using the knowledge of the known network attacks, classification-based model is created to determine whether or not the incoming event is anomalous. Lin *et al.* [14] proposed an intelligent intrusion detection algorithm to implement feature selection and decision rules generation by using SVM, Decision Tree, and Simulated Annealing. There are other researches relying on traffic-feature distribution and correlation [15], [19] for anomaly detection. For example, Yu *et al.* [10] found that

DDoS attack traffic is usually more similar to each other compared with flash crowds. Based on this phenomenon, a discrimination algorithm with the metric of flow correlation coefficient was proposed, which was applied to measure the similarity of suspicious flows, and to differentiate DDoS traffic from flash crowds. Besides, Li *et al.* [20] proposed a hybrid probabilistic metric to detect flooding attacks, which includes three models in the initial scheme, such as selecting suitable features, estimating their distributions, and defining proper correlations. The advantage of signature-based detection mechanism is that it can identify known anomaly with high accuracy, however, in the aspect of recognizing unknown or the variants of anomalous behavior, there are lots of false positives.

With the aim of solving the limitations of signature-based detection approach, a large number of profile-based approaches have been proposed. Due to the flexibility and importance to assist network management, the used techniques of these approaches vary in categories, such as statistical [21]–[24], clustering [16], [25], [26], and soft computing [11], [12]. For example, Ye and Chen [21] established chi-square theory for anomaly detection. In the work, a distance measure based on the chi-square test statistic was used to detect both a large departure of events from normal as anomalous and intrusions. Krügel *et al.* [22] proposed a statistical processing unit for detecting anomalous network traffic, more specifically, to detect the attacks that are rare such as R2L and U2R. Recently, principal component analysis (PCA) was used to analyze high dimensional network traffic dataset. Researches in [11], [13], [23], [27], [28] applied it to create a network profile called the digital signature of network segment using flow analysis (DSNSF) to predict the incoming network traffic. Besides, information-theoretic measures [16], [23], [27] were also used to create appropriate anomaly detection model, such as the shannon entropy, conditional entropy, relative entropy, rĺęnyi entropy and tsallis entropy.

Clustering refers to unsupervised learning algorithms which do not require pre-labeled data to extract rules for grouping similar data instances [29]. Faroughi and Javidan [28] proposed a new method called CANF which stands for clustering and anomaly detection by using nearest and farthest neighbors. In addition to that, Bigdeli *et al.* [25] proposed a two-layer cluster-based structure for anomaly detection. His experiment results demonstrate that the approach is fast, noise-resilient and incremental, and the false alarm rate is decreased from 5% to 15% compared to other alternatives. There are two assumptions in cluster-based techniques: on the one hand, normal data lies close to the nearest cluster centroid but anomalies are not; on the other hand, after obtaining different sizes of clusters, the instances in the smaller and sparser cluster are considered as anomalous and the thicker are considered as normal.

Recently, soft computing techniques including genetic algorithm, fuzzy logic and ant colony optimization are used for anomaly detection. Based on the behavior of biological processes, these algorithms can be employed to search optimal solution of the problem, especially in the network environment with most uncertainties and imprecisions. Hamamoto *et al.* [12] proposed a scheme with the combination of genetic algorithm and fuzzy logic. In his work, genetic algorithm was used to generate a digital signature, while the fuzzy logic was applied to decide whether an instance represents an anomaly or not.

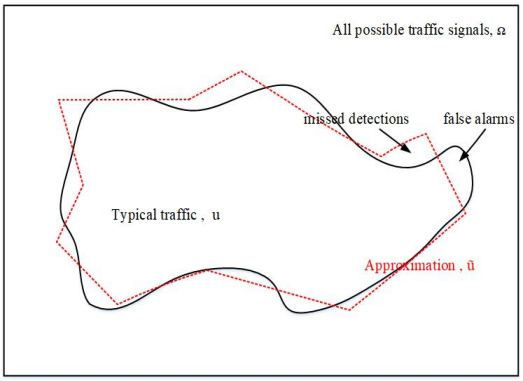


FIGURE 1. A simplified picture of anomaly detection [30].

After analyzing the related researches of profile-based approach, we find that it is critical to create a model that can fine-grained describe the traffic behavior, so that different anomalies arising simultaneously can be detected. We seek to construct a traffic profile and estimate the fittest of the incoming traffic behavior with the profile efficiently, as shown in Fig.1. However, the chose attributes of previous researches are single-dimension. Some of them are volume-based, while the others are spatial-based, which could not summarize the whole usage of network and all the behaviors of end-users. Besides, the real traffic pattern is changing over time, therefore, it is necessary to update the created profile to adjust to the dynamic traffic without much computation. Although, the work in [30] devised a basis evolution framework, much more computation is needed. Hence, this work extracts nine single-attribute and thirty-nine dual-attribute metrics from four dimensions ranging from temporal, spatial, category and intensity to profile the behavior of all IPs in the observed network. The main motivation of our work is to provide a comprehensive metrics set for traffic characterization and construct an adaptive high-efficiency baseline profile for anomalies detection.

III. THE PROPOSED APPROACH FOR ANOMALY DETECTION

The efficiency of profiled-based anomaly detection approach is fully related to traffic behavior characterization. Reference in [31] showed that the traffic behavior is currently composed of cycles and bursts, which either has periodic features due to the working days from Monday to Friday, or has a specific behavior pattern that occurs frequently in a specific period of time, e.g. access to the news or E-mails every morning.

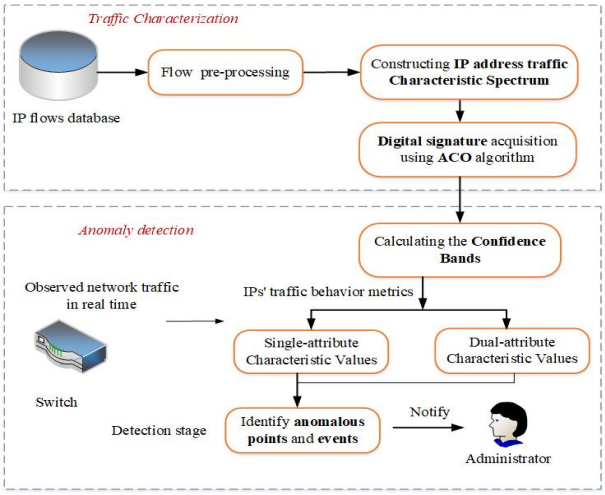


FIGURE 2. Modules of anomaly detection scheme.

In other words, the traffic behavior is directly affected by the user's behavior.

The idea of our work comes from the concept of behavior ethogram in animal ethology [32], which is a complete record of the common behavior of animals of the same species. Long-term observation, and an accurate and detailed record of their behaviors are required to establish it. Hence, it can be used to describe all the animals' behavior. Similarly, the IP traffic behavior characteristic spectrum is composed to profile the IPs' behavior in the observed network.

The proposal of the work is detailed in Fig.2. Our methodology is divided in two modules: (1) Traffic Characterization, (2) Anomaly Detection. The former forms the baseline behavioral framework of the normal traffic with the steps of constructing IP traffic behavior characteristic spectrum and generating digital signature matrix based on it. The latter performs anomalous-points detection and analysis, namely, anomaly clustering, so that a great many of unnecessary interventions are reduced.

A. IP FLOW PRE-PROCESSING

The IP flows used in this paper is collected from the border of China Education Research Network (CERNET) in Jiangsu Province, there may exist anomalous traffic if no filtering technique is applied. One important step of our research is to create an efficient basis traffic profile to depict normal traffic behavior. Hence, the collected real traffic is filtered by Network Behavior Observation System (NBOS) to obtain anomaly-free historical dataset.

NBOS is a network traffic behavior monitoring system for monitoring and managing CERNET's service quality and security status [33], including the identification of Internet Background Radiation traffic and the detection of DoS and DDoS. The former refers unsolicited one-way traffic, a type of unwanted traffic on the Internet, and the latter is an attack behavior. Not only of the DDoS traffic can cause inconvenience for users and revenue loss for service providers, but it can also flood a predetermined target through a series of malformed or malicious packets.

In addition to detecting malicious traffic behavior, NBOS can also recognize the application type of IP flow record by incorporating attributes of the source and destination ports, upper layer protocol, packet arrival interval and bidirectional packet ratio. After the filtration of real traffic, the rest traffic is applied to compose IP traffic behavior characteristic spectrum for traffic characterization.

B. COMPOSING IP TRAFFIC BEHAVIOR CHARACTERISTIC SPECTRUM

According to the concept of behavior ethogram in animal ethology, the definition of IP address traffic behavior characteristic spectrum is given, then the processes of composing it are detailed, including extracting IP address traffic behavior metrics, calculating IP address traffic behavior characteristics and stable characteristics sifting.

Definition 1: IP address traffic behavior characteristic spectrum is the full record of the stable characteristics obtained from the IP addresses during the observed cycle, which is used to profile the behavior of all IPs in the managed network, and provide data for the behavior description of each IP.

1) EXTRACTING IP ADDRESS TRAFFIC BEHAVIOR METRICS

In order to systematically and comprehensively profile the traffic behavior of the IPs, forty-eight metrics are extracted, including nine single-attribute and thirty-nine dual-attribute metrics from the dimensions of temporal, spatial, category and intensity, which correspond to the duration time, the peer address, the application type and the number of packets and bytes contained in the flow, respectively.

a: EXTRACTING SINGLE-ATTRIBUTE METRICS FROM IP FLOWS

Single-attribute metric refers to the metric involving only one dimension, although, there are more than one metric in a single dimension. For instance, Ndg and Scp are contained in spatial dimension, as shown in Table 1. In an IP network, IP flow records are often used to profile the traffic behavior of the IPs. A flow is defined as a unidirectional sequence of packets between a particular source-and-destination IP address pair. The attributes of our data include the source IP(Srcaddr), the destination IP(Dstaddr), the source port (Srcport), the destination port(Dstport), protocol, the start time (Stime), the during time(Gtime), the applcation type(Toa), the number of packets(Pkts), the number of bytes(Bytes), the number of packets sent from source to destination IP(l_pkts), the number of bytes sent from source to destination IP (l_bytes), the number of packets received from destination to source IP(r_pkts), the number of bytes received from destination to source IP(r_bytes) and etc..

The basic metric for temporal dimension is persistence, which is used to represent the duration time of each activity. As an IP flow record is considered as an independent

TABLE 1. Single-attribute metrics and their semantics.

Dimensions	Metrics	Semantics	Calculate Formula
Temporal	Drt	the duration of a flow	formula (1)
Spatial	Ndg	the number of peer addresses	formula (2)
	Scp	the number of peer address segments	formula (3)
Category	Dvs	the number of flow application types	formula (4)
	Ocr	the numbers of flows	formula (5)
Intensity	IstSendPkts	the number of packets sent of the flow	formula (6)
	IstSendByts	the number of bytes sent of the flow	Ref.formula (6)
	IstReceivePkts	the number of packets received of the flow	Ref.formula (6)
	IstReceiveByts	the number of bytes received of the flow	Ref.formula (6)

IP activity, the rule of time for the traffic behavior is described in terms of persistence. Its persistence (Drt) corresponds to the duration time of the flow.

$$Drt = Gtime \tag{1}$$

The basic metrics for spatial dimension of an activity are the number of locations (Ndg) that the activity occurs, and the space distribution of these locations (Scp). The occurrence times of an activity in a specific location refers to an association between it and the location. Ndg and Scp are used to depict the spatial behavior of the IP address. Note that the observed IP (source IP) in each IP flow record communicates with only one peer IP (destination IP), the Ndg is equal to the number of peer IP addresses it communicates with, and Scp is equal to the number of address blocks that the peer IP belongs to. The more address blocks associated with the observed IP, the wider the communication range is.

$$Ndg = card(DisIP) \tag{2}$$

$$Scp = card(DisSct) \tag{3}$$

Activity category dimension describes the number of behavioral categories of an activity, denoted as activity diversity (Dvs). In an IP network, Dvs is equal to the number of application types that an IP flow record belongs to. The application type of an IP flow record in this article originates from the Network Behavior Observation System. Assumes that P is the number of flows in an observed cycle.

$$Dvs = card\{Type_i | A_i \in P\} \tag{4}$$

$$Ocr = card(P) \tag{5}$$

$$IstSendPkts = \begin{cases} l\_pkts(Dstaddr = DisIP) \\ r\_pkts(Srcaddr = DisIP) \end{cases} \tag{6}$$

The intensity dimension describes the extent of the activity affects the system, expressed by the total number of occurrences of an activity (Ocr) and the intensity of each activity (Ist). The more packets and bytes sent or received, the greater influence on the network. There are four metrics used to describe the activity intensity of the IPs, including the number of packets sent (IstSendPkts), the number of bytes sent (IstSendByts), the number of packets received (IstReceivePkts) and the number of bytes received (IstReceiveByts). Taking IstSendPkts as an example, its intensity is equal to the number of packets sent from the observed IP to the peer IP address. In summary, the semantics and calculation methods of all single-attribute metrics are shown in Table 1.

b: EXTRACTING DUAL-ATTRIBUTE METRICS FROM IP FLOWS

Dual-attribute metric refers that there are two dimensions in a metric, in which one is divided based on the other, such as DvsInScp. This article refers the analogous method of animal ecology when constructing a dual-attribute metric set of IP activities. For instance, in the study of animal foraging behavior, an observed cycle is usually divided into N periods and the range of the animal's foraging behavior in each period is calculated [32], [34], [35]. Similarly, the IP activities are grouped into N subsets in terms of the temporal attribute, so the spatial range of each IP activity subset is counted. According to this approach, the dual-attribute metrics divided based on temporal, spatial and category dimensions are listed below.

DUAL-ATTRIBUTE METRICS DIVISION BASED ON TEMPORAL

This paper adopts two kinds of partition methods. (1) Dividing according to unit time. The IP flow with its start time fallen into the unit time is divided into a subset to describe the changes of other divided attributes. By using this division approach, the rate of changes of other attributes can be described. For example, the activity range per unit time is called the change rate of it, and is used to describe whether the activity range is changing rapidly or not. (2) Dividing by the period of time. This division approach is applied to characterize the rhythm behavior of the IP. In an IP network, the rhythm behavior of an IP address refers to the periodicity behavior reflected by the traffic, e.g. whether the IP is active during the daytime or at night, or the behavior often appears in a specific period of time. In this paper, one day is divided into six periods. Metrics in each period are counted with their semantics as shown in Table 2.

DUAL-ATTRIBUTE METRICS DIVISION BASED ON SPATIAL

There are two division methods considered: (1) dividing the flow based on peer IP address. This approach can be used to depict the degree of association of the observed IP. (2) Dividing the flow based on peer IP address segments that indicates the size of the communication range of the observed IP. The more address blocks associated with the

TABLE 2. Dual-attribute metrics division based on temporal and their semantics.

Dividing Method	Dimensions	Metrics	Semantics	Calculate Formula
Divided by unit time	Spatial	NdgPerUnit	the number of peer IPs per unit time	(2)
		ScpPerUnit	the number of peer IP segments per unit time	(3)
	Category	DvsPerUnit	the number of application types the flow belongs per unit time	(4)
		OcrPerUnit	the number of flows per unit time	(5)
	Intensity	IspPerUnit	the number of packets sent per unit time	(6)
		IsbPerUnit	the number of bytes sent per unit time	Ref.(6)
		IrpPerUnit	the number of packets received per unit time	Ref.(6)
		IrbsPerUnit	the number of bytes received per unit time	Ref.(6)
Divided by time period	Temporal	DrtInRtm	the duration of the flow in each period of time	(1)
	Spatial	NdgInRtm	the number of peer IPs in each period of time	(2)
		ScpInRtm	the number of peer address segments in each period of time	(3)
	Category	DvsInRtm	the number of application types in each period of time	(4)
		OcrInRtm	the number of flows in each period of time	(5)
	Intensity	IspInRtm	the number of packets sent in each period of time	(6)
		IsbInRtm	the number of bytes sent in each period of time	Ref.(6)
		IrpInRtm	the number of packets received in each period of time	Ref.(6)
		IrbInRtm	the number of bytes received in each period of time	Ref.(6)

TABLE 3. Dual-attribute metrics division based on spatial and their semantics.

Dividing Method	Dimensions	Metrics	Semantics	Calculate Formula
Divided by peer IP	Temporal	DrtPerNdg	the duration of the observed IP communicate with each peer IP	(1)
	Category	DvsPerNdg	the number of application types of each peer IP	(4)
		OcrPerNdg	the number of flows of the observed IP communicate with each peer IP	(5)
	Intensity	IspPerNdg	the number of packets sent between the observed IP and the peer IP	(6)
		IsbPerNdg	the number of bytes sent between the observed IP and the peer IP	Ref.(6)
		IrpPerNdg	the number of packets received between the observed IP and the peer IP	Ref.(6)
		IrbPerNdg	the number of bytes received between the observed IP and the peer IP	Ref.(6)
Divided by peer IP segments	Temporal	DrtInScp	the duration of the observed IP communicate with each peer IP segment	(1)
	Category	DvsInScp	the number of application types of each peer IP segment	(4)
		OcrInScp	the number of flows of the observed IP communicate with each peer IP segment	(5)
	Intensity	IspInScp	the number of packets sent between the observed IP and each peer IP segment	(6)
		IsbInScp	the number of bytes sent between the observed IP and each peer IP segment	Ref.(6)
		IrpInScp	the number of packets received between the observed IP and each peer IP segment	Ref.(6)
		IrbInScp	the number of bytes received between the observed IP and each peer IP segment	Ref.(6)

observed IP, the wider the communication range is. The metrics and semantics are given in Table 3.

DUAL-ATTRIBUTE METRICS DIVISION BASED ON CATEGORY

This division method puts each type of application into a subset and constructs the category behavior matrix to depict their behaviors. The statistic of the temporal dimension attributes is used to describe the duration time of each type of application; the statistic of the spatial dimension attributes is used to describe the number of peer IPs and the peer IP segments that occur in each type of application, and the statistic of the intensity dimension attributes is applied to describe the number of flows of a specific type and the intensity of each type of application. The metrics and semantics are given in Table 4.

2) CALCULATING IP ADDRESS TRAFFIC BEHAVIOR CHARACTERISTICS

Definition 2: IP address traffic behavior characteristic refers to the abstraction of the different behavioral trait of the IP address during their communication, which is specifically represented by the form of the value for each metric.

Four types of numerical value form are obtained after the analysis of nine single-attribute metrics and thirty-nine dual-attribute metrics. (1) Single variable. It is described directly by the value of the metric itself and its characteristic value is the metric value. (2) Dimension-determined vector.

TABLE 4. Dual-attribute metrics division based on category and their semantics.

Dividing Method	Dimensions	Metrics	Semantics	Calculate Formula
Divided by category	Temporal	DrtPerClss	the duration of each type of flow	(1)
	Spatial	NdgPerClss	the number of peer IPs of each type of flow	(2)
		ScpPerClss	the number of peer IP segments of each type of flow	(3)
		OcrPerClss	the number of flows of each type of flow	(5)
	Intensity	IspPerClss	the number of packets sent of each type of flow	(6)
		IsbPerClss	the number of bytes sent of each type of flow	Ref.(6)
		IrpPerClss	the number of packets received of each type of flow	Ref.(6)
		IrbPerClss	the number of packets received of each type of flow	Ref.(6)

A set of values are used to describe the characteristics of the IP traffic behavior. (3) Size-uncertain set. The statistic values of the mean and variance are used to summarize the overall characteristic value. (4) Complex value. For this form of numerical values, it requires case by case according to the objective in different research areas.

a: CALCULATING THE BEHAVIOR CHARACTERISTICS FOR SINGLE-ATTRIBUTE METRICS

Two kinds of numerical value form appear in single-attribute metrics, including single variable and size-uncertain set. The metrics value of single variable is calculated only once during the entire observed cycle, such as Ndg, Dvs and Scp, with the numerical value of the each metric as their characteristic value. But, the size-uncertain set is calculated multiple times throughout the observed cycle with the same meaning of each calculation, such as IstSendPkts and IstSendByts. Mean and variance are used to summarize the meaning of the metric.

b: CALCULATING THE BEHAVIOR CHARACTERISTICS FOR DUAL-ATTRIBUTE METRICS

The numerical value form of each dual-attribute metric is affected by different division methods. The approach of dividing all flows into a size-fixed flow subsets according to the rhythmicity and the flow application types is used. In order to describe the rhythmicity behavior of each IP, one day is divided into six periods, and the flows with their start time fallen in each period are put correspondingly. In addition, IP flows of different application types are divided into different flow subsets and the size of subset is determined by the number of application types. On one hand, when the single-attribute metric value of each flow subset is in the form of single variable, the corresponding dual-attribute metric value, in this case, is a dimension-determined vector, whose characteristic value is the numerical value of each position in the vector, such as NdgInRtm, ScpInRtm, NdgPerClss, OcrPerClss, DvsInRtm, OcrInRtm and ScpPerClss. On the other hand, when the single-attribute metric value of each flow subset is in the form of size-uncertain set, the corresponding dual-attribute metric value, in this case, takes the form of complex value, expressed as a dimension-determined vector set. The characteristic value is the statistical value of each vector in the vector set, such as IspInRtm, IrpInRtm, IspPerClss, IsbPerClss, IsbInRtm, IrbInRtm, IrpPerClss and IrbPerClss.

3) THE STABLE CHARACTERISTICS SIFTING

The stability of IP address traffic behavior characteristics is divided into existence stability and association stability. The existence stability of the behavior characteristics refers to the long-term nature of the managed network, regardless of the characteristics occurring on the specific IP or not. The stability characteristics in a cycle reflect the steady traffic behavior of IPs, which can summarize the behavior of all IP traffic in the managed network. The association stability of the behavior reflects whether the characteristic is associated to the IP for a long time or not. Association stability is measured by using association rate, which is equal to the ratio of the associated duration time to the total active length of IP address. A low association rate indicates that the stability of the association between the characteristic and the IP address is not strong.

In the paper, existence stability is used for stable characteristics sifting with the occurrence frequency of the characteristic as the stability factor. The characteristics with low frequency are eliminated when constructing the characteristic spectrum. In order to determine the appropriate frequency threshold, one week are chose as the observed cycle. After observing for three weeks continuously, the feature-frequency distribution chart is drawn, as shown in Fig. 3. Although, there are some differences of the number of features in the features-frequency chart among the three cycles, they all show obvious heavy-tailed trait. Most of the traffic characteristics are located at the head of the curve, with the frequency less than 0.02 and the stability is weak. In the experiment, 2% is selected as the threshold of stable occurrence frequency. The number of stable characteristics of the three cycles are 1099, 1113, and 1073, respectively.

C. CONSTRUCTING DIGITAL SIGNATURE MATRIX

Ant colony optimization algorithm simulates the behaviors of real ant colonies to find the shortest route between a food source and their nest. The basic principle of it is swarm intelligence, defined as a population of concurrent and globally asynchronous agents that cooperate to find solutions to complex optimization problems [36].

Ants use statistics and probabilities to travel through the search space and make convergence of the entire colony. Although, each ant with the ability of having a feasible solution, the cooperation of the individuals can achieve a highest-quality solution. ACO has been applied in many

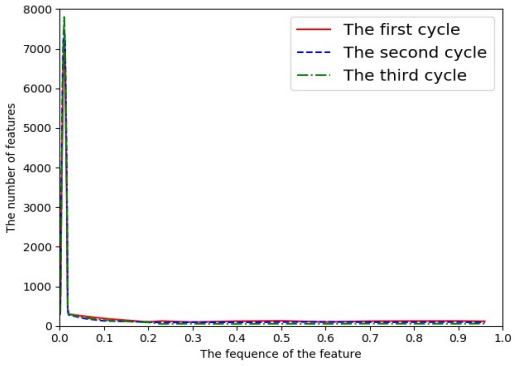


FIGURE 3. Feature-frequency distribution chart of three cycles.

research areas with outperforming performance based on the traits of distributed collaboration, self-organizing and positive feedback. For instance, fault localization [37], efficient energy management in wireless sensor networks [38], route discovery and network reconfiguration [39] are all found in communication networks.

In this paper, ACO algorithm is used to capture the digital signatures of the normal traffic behavior profile. Clustering technique is used by ACO to obtain digital signatures with three operations. (1) Build solutions. It assigns movements to the concurrently and asynchronously ants according to the states of problem. (2) Local Search. It is used to test and evaluate the solutions, in the meantime, remove unpromising solutions.(3) Pheromone Update. It can help the ants to seek new paths, so that a near-optimal solution is acquired with the pheromone's intensity increased or decreased.

The search space is represented as a graph G (V, B), where V is the nodes set and B is the set of edges. In our work the node is the characteristic value of each metric. Assumes that the characteristic values represent the elements to be clustered, while edges connect the each of them to the centroid of the groups. As every cluster has its centroid, which is a new node located in the middle of the cluster. The purpose of our work is to find these centroid, which is called digital signature and use each of them on behalf of the cluster. The objective function in Eq.(7) is applied for this purpose, when choosing five clusters, we get the best objective function.

$$J = \sum_{i=1}^{E} \sum_{j=1}^{k} \sqrt{\sum_{a=1}^{A} (x_{ia} - c_{ja})^2} \tag{7}$$

The metrics set is composed of each element i and will be grouped to the cluster j, in which $j = 1, ..., K$. Variable E represents the quantity of metrics to be clustered, while A indicates data dimensionality, namely, the number of characteristic values to be processed of each metric. The collected elements include four cycles with each week as one cycle. Variable x_{ia} denotes the characteristic value of each metric, while c_{ja} stores the cluster center value j of each cluster. The first and second sums apply the aforementioned comparisons between all K cluster centroid and E elements, and the third sum is related to the Euclidean distances between x_{ia} and c_{ja}.

The output of J corresponds to the minimized objective function of our clustering approach. Algorithm 1 below shows the pseudo code of calculating the digital signature matrix.

Algorithm 1 Pseudo Code of Calculating the Digital Signature Matrix

Input: The characteristic values $(f_1, f_2, \ldots, f_n)$ of each metric, N: the number of metrics(48),K:the number of clusters(5).

Output: $\boldsymbol{M}$: Digital signature matrix.

```
for i = 1 to N do
    while Stopping condition is not reached do
        Create Solution
        Evaluate solutions through the objective function
        Update the pheromone trail
    end while
    for j = 1 to K do
        Calculate centroid of each cluster of the best solution found
        Pushback the centroid into a vector V
    end for
    Pushback the vector V into M
end for
return M
```

D. ANOMALY DETECTION AND CHARACTERISTIC SPECTRUM UPDATING

The digital signature matrix based on IP traffic behavior characteristic spectrum is used as the basic evaluation framework of the network segment, as shown in Fig.4. Any deviations slight from the digital signatures are considered as anomalies, hence, hundreds of alarms would be generated. In order to reduce the number of false positives and unwanted interventions, confidence bands are used to detect anomalous point. Then, a simple cluster algorithm is applied to analyze these anomalous points with the aim of attributing them to a single anomaly event. At last, characteristic spectrum updating strategy is given to adjust the changes of the dynamic traffic.

1) ANOMALOUS-POINT DETECTION AND ANALYSIS

The upper and lower limits are used for anomalous point's identification, any points of p that fall outside these thresholds are declared to be anomalous.

$$EL_{up} = \mu_i + q \times \sigma_i \tag{8}$$

$$EL_{down} = \mu_i - q \times \sigma_i \tag{9}$$

where EL_{up} is the upper threshold, EL_{down} is the bottom threshold, μ_i is the digital signature for each cluster of one metric, σ_i is the standard deviation of each cluster. The label q is equal to the cutoff value (Γ) of the total anomaly score from the Precision-Recall curve. Γ is defined by argmax (Precision + Recall), which value is obtained by shifting and calculating the associated precision and recall.

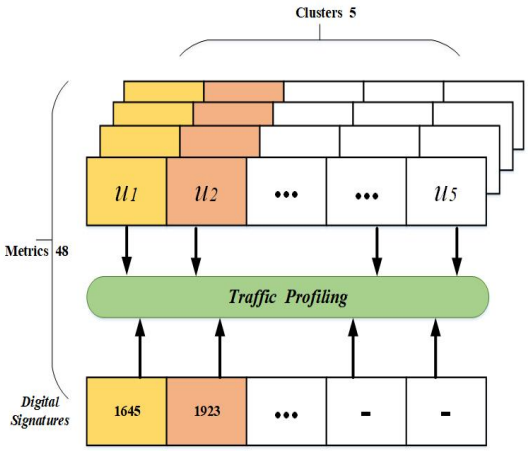


FIGURE 4. Digital signature matrix generated using historical traffic behavior.

Once the anomalous data points are recognized, a bottom-up hierarchical clustering algorithm is used to group these anomalous points. The assumption is that for most anomaly types, the time of the anomalous points caused by the same event are close each other. Based on this assumption, Euclidean distance is calculated as the similarity standard in our cluster algorithm. Any anomalous points occur within small time gaps are tend to group in a cluster, therefore our approach has the capability to identify short-time anomalous events. Algorithm 2 below gives the pseudo code of identifying anomalous points and anomalous events. Our cluster approach is generally and can be used in other detectors, as not any specific properties are considered.

2) CHARACTERISTIC SPECTRUM UPDATING TO ADJUST THE BASELINE PROFILE

As the applications and users in the network changing constantly, IP traffic behavior will also change, which will lead to the emergence of new behaviors or characteristics. Besides, some stable characteristics may no longer appear, therefore, updating the IP traffic behavior characteristic spectrum is necessary. According to the degree of variation in the traffic, the updating process includes the following operations.

(1) Replacement of characteristics. When the network traffic is slightly fluctuating, or there are few IPs' traffic behavior has changed, in this case, only some characteristics in the characteristic spectrum are no longer stable. Hence, the stability of these characteristics needs to be calculated and replaced in the characteristic spectrum.

(2) Reconstruction of the characteristic spectrum. When the network structure changed, or the DDoS attack happened, in this case, significant changes of the traffic behavior in the managed network will occur. Therefore, the characteristics in the characteristic spectrum are likely no longer stable. In this case, the previous features within the spectrum is not applicable and the reconstruction of the characteristic spectrum is required.

Algorithm 2 Pseudo Code For Anomalies Detection

Input: The characteristics (f_1, f_2,…,f_{48}) of each real time flow, Digital signature matrix:$\boldsymbol{M}$, and the corresponding threshold of each digital signature: EL_{up}, EL_{down};

Output: $\boldsymbol{E}$: Anomalous events, namely, clusters of anomalous-points $C_i(p_1,p_2\ldots p_n)$;

1: **for** f_i in each flow **do**
2: Judging that each characteristic f_i whether or not in the cluster that the corresponding digital signature belongs to
3: Judging that f_i whether or not in the range of between EL_{up} and EL_{down}
4: If there are more than six characteristics of an IP flow out of the range of the corresponding confidence bands, considering the flow as anomalous
5: Put the flow as anomalous-point p in the anomalous database D
6: **end for**
7: **repeat**
8: Computing all pair-wise similarity by using the attribute of start time within the flows $p \in$D
9: Finding two clusters that nearest each other
10: Merging them together to form a new cluster C
11: Computing the similarity distance from C to all other clusters
12: **until** there is only one cluster left
13: **return** $\boldsymbol{E}$

The rate of change for the characteristics within characteristic spectrum is calculated to determine whether the characteristic spectrum requires to be reconstructed or not. The characteristics in the initial characteristic spectrum called initial characteristic. After the observed cycle ends, it is necessary to proceed characteristics alternation. Then, the ratio of the initial characteristic in the current characteristic spectrum is calculated and called characteristic maintenance rate. Other characteristics, rather than initial characteristics account for in the current characteristic spectrum called the rate of change. When the rate of change is sufficiently large (the threshold of the rate of change is 5% in the experiment), in this case, the characteristic maintenance rate is small, which indicates most of the IP traffic behaviors have changed. It is necessary to reconstruct the characteristic spectrum.

IV. EVALUATION

All experiments in this article are performed on a 2-way Intel Xeon server with one Intel(R) Xeon(R) CPU E5-2650 processor on each path. Each processor contains 8 cores at a frequency of 2.00 GHz, with the memory of 128GB. The algorithm is implemented in C++ and python language. The proposed approach for anomaly detection is evaluated with real data of China Education Research Network backbone

TABLE 5. Anomalies injected in 9/28/2018.

Anomalies	Start time	End time	The number of flows
Port Scan	12:00	16:00	3280
DDoS	12:00	16:00	5280
Flash crowd	12:00	16:00	10760

and synthetic data. In the experiment, traffic characterization phase is evaluated first. Then, the ability and the performance of our anomaly detection algorithm are evaluated. Finally, the complexity of the scheme and the comparisons of our algorithm with other similar alternatives are evaluated.

A. TRAFFIC CHARACTERIZATION EVALUATION

With the purpose of verifying the proposed approach whether can operate in a real network environment or not, the NetFlow data collected from some border routers of China Education Research Network backbone in Jiangsu province is used. NetFlow is a popular traffic monitoring tools in the Internet nowadays, due to more information provided than SNMP, such as transport protocol, source and destination IP, source and destination ports and etc.. As large volume traffic in backbone, the sampling rate 1:256 is used to implement the collection protocol. The data was collected from August 20 to October 1, 2018, about five weeks, including three institutions, and a total of "67,584" educational network IPs.

The dataset is separated into two groups, four consecutive weeks' cleaned traffic is applied to create the digital signature matrix, and the fifth week is used for anomaly detection. As the references of [1] and [40] said, there are two broad anomalous categories that affect network system, performance related anomalies and security-related anomalies. Hence, anomalous traces such as Port Scan, Distributed Denial of Service (DDoS) and Flash Crowd collected and compounded with the cleaned data are applied to verify the capability of identifying different anomalous traffic arising simultaneously in the wild. Table 5 presents the time periods in which these anomalies are injected.

In order to profile the behavior of all IPs in the observed network, historical dataset of one institution lasts for four consecutive weeks is used to compose IP address characteristic spectrum. Assumes that network behavior characteristics reflect the user's behavior habits, in order to verify this assumption, characteristic spectrum with the temporal division metrics is given due to the space limitation. Six periods are separated of one day. We observe four consecutive weeks, and find that the traffic behavior of this period is basically stable with smaller fluctuations. Average value is calculated as the final results, as shown in Fig.5.

As Fig.5 shows, the volume of the traffic is large and with wider communication range between the afternoon and midnight. However, it is relatively small at other period of time. This phenomenon can be explained that most users in the campus are attending classes during the daytime, while other absentees contribute to relatively small amount of traffic. Furthermore, online learning or entertainment become more in

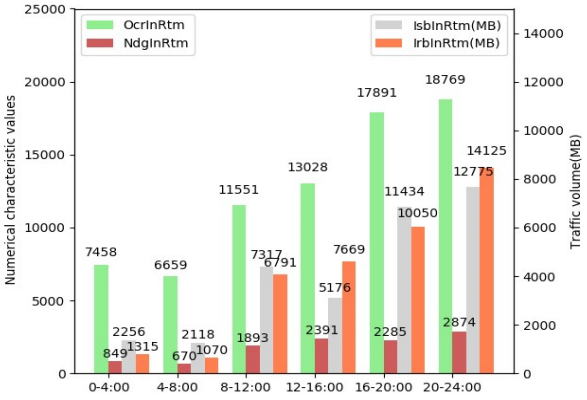
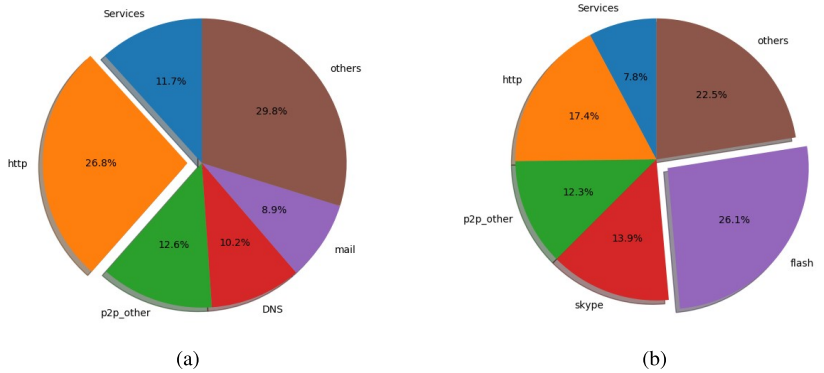


FIGURE 5. Traffic behavioral characteristic spectrum based on temporal division method.

the evening, and contribute to relatively huge traffic. In order to verify our inference, the percentage of specific applications of the flows between 8:00-12:00 and 16:00-22:00 are calculated. A total of 15 different application types are found. The percentage of each different application are shown in Fig. 6. As it shows, the traffic of http, p2p_other, DNS, etc. are relatively large at 8:00 to 12:00(Fig. 6a). It can be inferred that users browse news, blogs, mails, and other activities during this period of time, which is more conformed to the behaviors of the staff in the campus. However, the traffic of http, flash, skype application at 16:00 to 22:00 is relatively large (Fig. 6b). It demonstrates that users in this period are engaged in web services, video, social activities and other entertainment activities, which are closer to the students' behavior.

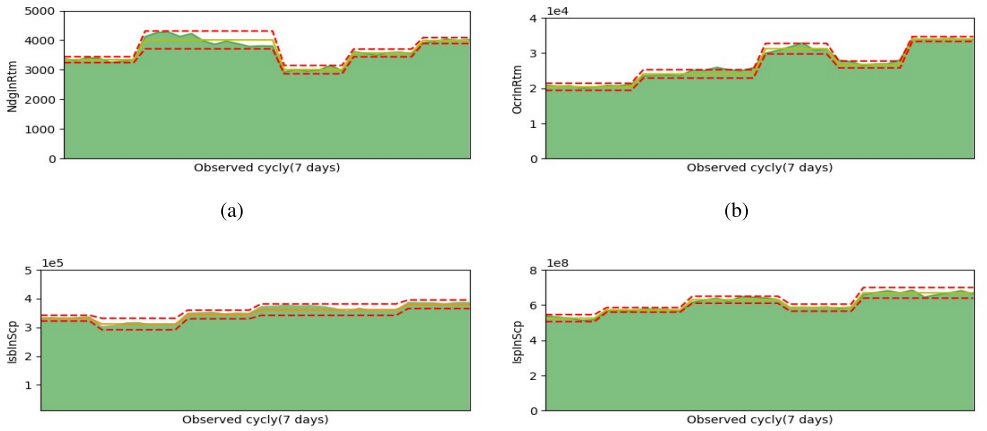
After profiling the behavior of the observed network, digital signature matrix is constructed by using ACO for anomaly detection. Due to the space limitation, only the digital signatures of dual-attribute metrics division based on temporal and spatial are given to verify the feasibility of our approach, including four metrics, NgdInRtm, OcrInRtm, IspInScp, IsbInScp. Fig.7 shows the behavior of real measurements and their digital signatures, where the solid orange lines represent the digital signatures generated by ACO, while the green area and the two red dotted line represent the real traffic behavior characteristic values and the confidence bands, respectively. As the figure shows, the digital signatures of the four metrics could effectively on behalf of the normal traffic behavior cluster of the observed network. Therefore, any characteristics of the IP traffic behavior out of the confidence bands are considered as anomalous.

The accuracy of the digital signatures for each metric is measured by adopting Normalized Mean Square Error (NMSE). NMSE was proposed by [41]in 1993, used to evaluate the difference between the expected and what is actually measured by calculating the normalized difference between them. If the value is equal to zero, it indicates the expected is exactly equal to the actually measured, otherwise, it indicates a large distance between them. Our work uses 48 metrics



(a) (b)

FIGURE 6. The percentage of different applications, (a) 8:00-12:00 o'clock. (b) 16:00-22:00 o'clock.



(a) (b) (c) (d)

FIGURE 7. The behavior of real measurements and their digital signatures, (a) NgdInRtm. (b) OcrInRtm. (c) IspInScp. (d) IsbInScp.

for traffic characterization and anomaly detection, we don't give the NMSE values for all the metrics. In order to verify the feasibility of our approach, eight metrics are random selected, including two single-attribute metrics and six dual-attribute metrics. As observed in Fig.8, the proposal shows good results for these metrics with the error indices closer to zero (bellow 0.1). The NMSE is calculated in Eq.10.

$$NMSE = \frac{N * \sum_i (x_i - y_i)^2}{(\sum_i x_i * \sum_i y_i)} \tag{10}$$

where N is the number of observed samples, x_i and y_i are the observed and the expected values for the sample i, respectively.

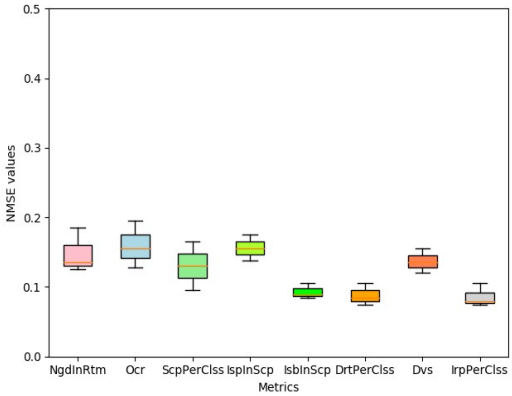


FIGURE 8. The average NMSE of each metrics for four cycles.

B. ANOMALY DETECTION EVALUATION

To properly evaluate the anomaly detection system proposed in the work, anomalous events of DDoS, Flash Crowd and Port Scan detected by NBOS system are used to test the proposed anomaly detection mechanism.

Port Scan anomaly related to a single source IP address transmits TCP connection messages to a single destination IP, generally, from a specific source port to a wide range of destination ports. DDoS is composed of multi sources, in which, populations of compromised computers are controlled by the command and control sever, so that the attacker can set up his attack against the targeted service at his will.

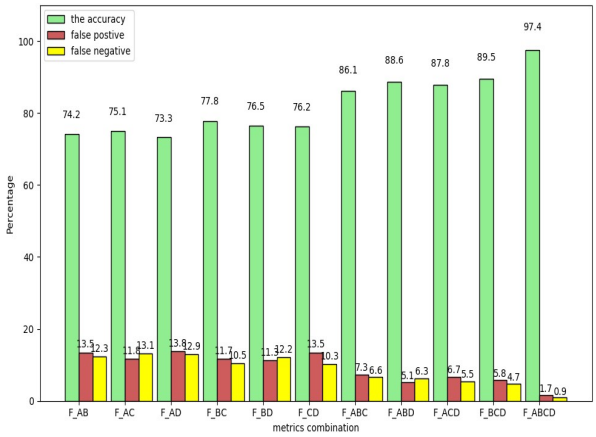


FIGURE 9. Identifying the anomalous with different metric combinations.

Flash Crowd is defined as large floods traffic occurring when rapid growth users access to the network resource, e.g. the traffic in double-eleven shopping of ''www.taobao.com''. Unlike Port Scan and DDoS attack, legitimate traffic is consisted in it. However, if there is no enough time reacts to those flash events, they can seriously flood or lead to a complete failure to web service. In our work, these anomalies are in combination with real traffic to form the real network scenario containing both volume-based and spatial-based anomalies.

Three metrics are used to evaluate the performance of the algorithm, such as the accuracy rate, false-positive rate, and false-negative rate. Single-attribute and dual-attribute metrics are incorporated to test the efficiency of different metrics combination, where ''A'' is representative of single-attribute metrics for short, ''B'' is representative of dual-attribute metrics division based on temporal for short, ''C'' is representative of dual-attribute metrics division based on spatial for short, and ''D'' is representative of dual-attribute metrics division based on category for short, respectively. The accuracy rate, false-positive rate, and false-negative rate of the artificial dataset under different metrics combination are evaluated as shown in Fig.9. Experimental results demonstrate that the combination of single-attribute and dual-attribute metrics sets are better than the other combinations.

Then, we testify our scheme could detect different anomalies that arise simultaneously in a real network scenario. Two types of different metrics combination have been random selected for three times to verify our approach with the average value as the final result. If in this case, it can verify our algorithm, use four types of different metrics combination could also work in the real network scenario with high accuracy. The top-3 information(source and destination IP addresses and Ports) related to these anomalies occurring for a time interval are shown by Fig.10. Experimental results demonstrate that our scheme can detect different anomalies that arise simultaneously in the wild in the worst cases.

C. COMPLEXITY ANALYSIS

Three key components are noted in our approach, including IP flow pre-processing, composing IP address traffic behavior characteristic spectrum for constructing digital signature matrix, anomaly detection and updating IP address traffic behavior characteristic spectrum.

In NBOS system, DoS and DDoS attacks are detected by using information theory technology. Entropy is the most common metric for qualitative analysis IP flow information. With the number of flow entries (e) processed in the analyzed time interval, $O(e^2)$ can be obtained.

The complexity of this work is mainly related to traffic characterization phase, since lots of computations are required in this stage, such as the operations of composing IP address traffic behavior spectrum and the creation of digital signature matrix. If we spilt l intervals and use k predefined centers with the number of N characteristics, its computational complexity is $O(lkN)$. When choosing the population of m ants to search the best centers of the data, a quadratic complexity will be obtained $O(lkNm^2)$. However, the processes of composing IP address traffic behavior spectrum and anomaly detection require a normalization step with the computational complexity $O(n)$, respectively. Other calculations are constant in time, such as updating IP address traffic behavior spectrum. As the algorithm extracts (f) metrics for profiling the behavior of a specific network segment, thus the total complexity of the work is $O(fne^2 + nlkNm^2 + fn)$ with n days as an observed cycle. In the worst cases, if the number of iterations I as stop condition of the algorithm, the final complexity is $O(nlkNIm^2)$.

D. COMPARISONS WITH OTHER SIMILAR ALTERNATIVES

The proposed method is compared with other four similar alternatives, such as ACODS [2], outlier detection [24], Anomaly Detection System based on Genetic Algorithm and Fuzzy Logic [12], using GA and Fuzzy Logic for short, and the BasisEvolution framework [30]. The comprehensive and fine-grained behavior description of the work is compared with ACODS and Anomaly Detection System based on Genetic Algorithm and Fuzzy Logic first, and then, the performance is compared as shown in Fig.11.

The reference in [2] applied the volume-based attributes such as the number of flows, packets and bits to describe the traffic behavior from the perspective of traffic intensity only. If traffic analysis simply focuses on the heavy traffic, low-volume anomalous patterns could be missed. Although, some metrics such as traffic distribution in time and space are similar to ours, they are not as detailed as ours. For example, [12] analyzed the traffic behavior of the IP address from the perspective of spatial and intensity by using attributes of IP address, ports, the number of flows, packets and bits. These attributes are similar to our single-attribute metrics, but, is not detail or comprehensive than ours. In our work, some dual-attribute metrics are divided based on temporal, which could be used to describe the rhythmicity behavior of an IP activity. Similarly, the traffic distribution in space dimension is also

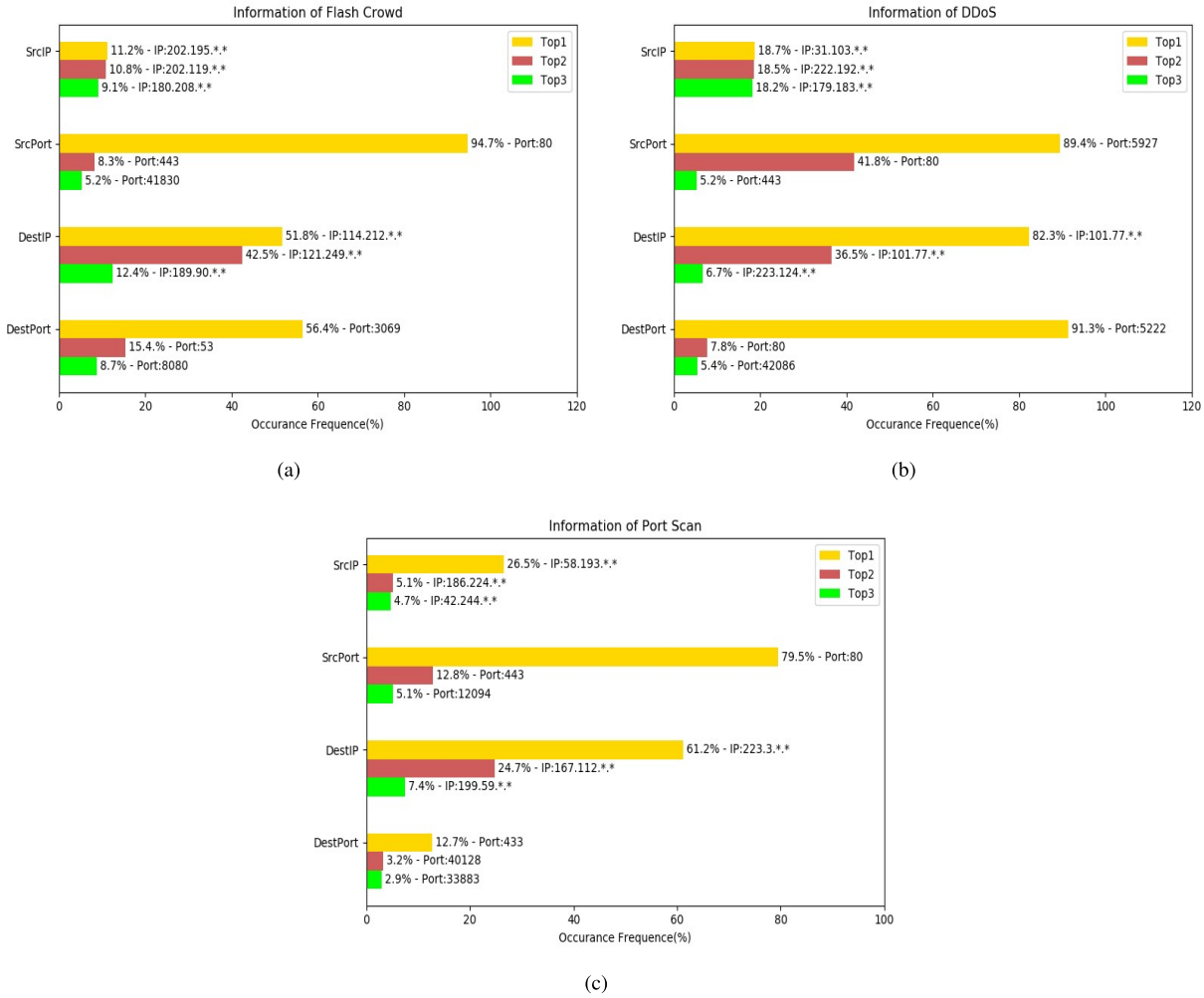


FIGURE 10. Different anomalies of the Top-3 information, (a) FlashCrowd. (b) DDoS. (c) PortScan.

considered, hence, not only the observed IP communicate with which end hosts, but their communication range can also be found. Results demonstrate that our work could fine-grained and comprehensive describe the traffic behavior than others.

Besides, our work achieve better performance than other similar alternatives, as shown in Fig.11. Although, ACODS and Anomaly Detection System based on Genetic Algorithm and Fuzzy Logic have similar accuracy compared with our approach, the F-measure of our work is the best among all the compared algorithm. What is more, the memory consumption of our work is small than others, as the matrix of our work is 48*5, while the references of [2] and [12] are 1440*3 and 1440*7, respectively. In a word, although the algorithm cannot completely replace other detection algorithms, it can be used as a supplementary to other detection mechanism to some extent. Therefore, the presented approach proves to be a valuable tool to assist in traffic analysis and network management due to the accurate detection ability, and the fine-grained and comprehensive traffic characterization ability.

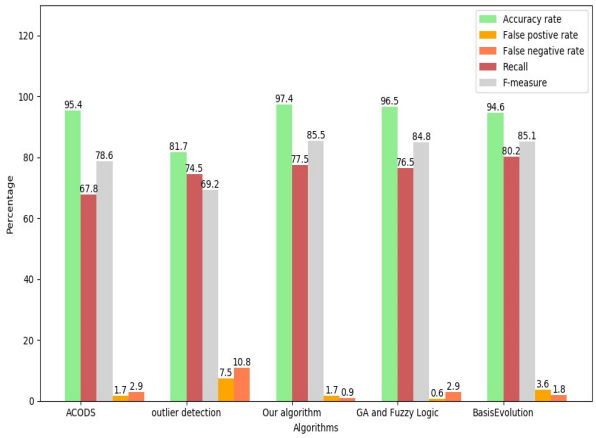


FIGURE 11. Comparisons of different algorithm.

V. CONCLUSION

In this paper, an adaptive profile-based anomaly detection system is proposed to help network management. The major contribution of our work is the application of IP address

traffic behavior characteristic spectrum for traffic characterization. Based on it, digital signature matrix is obtained as the baseline profile for anomaly detection. Regarding traffic characterization module, nine single-attribute and thirty-nine dual-attribute metrics are extracted to construct IP address traffic characteristic spectrum from the dimensions of temporal, spatial, category and intensity, which can be applied to discover the behavior information of IP address and provide data for the generation of digital signatures in a large number of noise-containing flow data. After the analysis of the traffic behavior of the institution (202.194.*.*/20), we find that the traffic has obvious rhythmic behavior, which conforms to the habits of different users in the network. In the detection and identification module, confidence bands and a general cluster technique are applied. Experimental results demonstrate that our detection algorithm can identify different anomalies that arise simultaneously in the wild with a high detection accuracy rate (97.4%) and a low false negative rate (0.9%). What is more, after cluster analysis of these identified anomalous points, the number of unwanted notifications to the administrator are decreased to a large extent. Our approach is autonomous and adaptive by adjusting the baseline profile with constant time computational complexity, therefore, the phenomenon of traffic drift cannot affect its accuracy.

ACKNOWLEDGMENT

Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of those sponsors.

REFERENCES

[1] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "Network anomaly detection: Methods, systems and tools," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 303–336, 1st Quart., 2014.

[2] G. Fernandes, Jr., J. J. P. C. Rodrigues, and M. L. Proença, Jr., "Autonomous profile-based anomaly detection system using principal component analysis and flow analysis," *Appl. Soft Comput.*, vol. 34, pp. 513–525, Sep. 2015.

[3] G. Thatte, U. Mitra, and J. Heidemann, "Parametric methods for anomaly detection in aggregate traffic," *IEEE/ACM Trans. Netw.*, vol. 19, no. 2, pp. 512–525, Apr. 2011.

[4] G. Nychis, "An empirical evaluation of entropy-based anomaly detection," in *Proc. Internet Meas. Conf.*, 2008, pp. 151–156.

[5] I. Hareesh, S. Prasanna, M. Vijayalakshmi, and S. M. Shalinie, "Anomaly detection system based on analysis of packet header and payload histograms," in *Proc. Int. Conf. Recent Trends Inf. Technol.*, Jun. 2011, pp. 412–416.

[6] X. Ma and Y. Chen, "DDoS detection method based on chaos analysis of network traffic entropy," *IEEE Commun. Lett.*, vol. 18, no. 1, pp. 114–117, Jan. 2014.

[7] A. Ziviani, A. T. A. Gomes, M. L. Monsores, and P. S. S. Rodrigues, "Network anomaly detection using nonextensive entropy," *IEEE Commun. Lett.*, vol. 11, no. 12, pp. 1034–1036, Dec. 2007.

[8] A. Patcha and J.-M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," *Comput. Netw.*, vol. 51, no. 12, pp. 3448–3470, Aug. 2007.

[9] A. A. Amaral, L. de Souza Mendes, B. B. Zarpelão, and M. L. P. Junior, "Deep IP flow inspection to detect beyond network anomalies," *Comput. Commun.*, vol. 98, pp. 80–96, Jan. 2016.

[10] S. Yu, W. Zhou, W. Jia, S. Guo, Y. Xiang, and F. Tang, "Discriminating ddos attacks from flash crowds using flow correlation coefficient," *IEEE Trans. Parallel Distrib. Syst.*, vol. 23, no. 6, pp. 1073–1080, Jun. 2012.

[11] G. Fernandes, Jr., L. F. Carvalho, J. J. P. C. Rodrigues, and M. L. Proença, Jr., "Network anomaly detection using IP flows with principal component analysis and ant colony optimization," *J. Netw. Comput. Appl.*, vol. 64, pp. 1–11, Apr. 2016.

[12] A. H. Hamamoto, L. F. Carvalho, L. D. H. Sampaio, T. Abrão, and M. L. Proença, Jr., "Network anomaly detection system using genetic algorithm and fuzzy logic," *Expert Syst. Appl.*, vol. 92, pp. 390–402, Feb. 2018.

[13] M. V. O. de Assis, J. J. Rodrigues, and M. L. Proença, Jr., "A seven-dimensional flow analysis to help autonomous network management," *Inf. Sci.*, vol. 278, pp. 900–913, Sep. 2014.

[14] S.-W. Lin, K.-C. Ying, C.-Y. Lee, and Z.-J. Lee, "An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection," *Appl. Soft Comput.*, vol. 12, no. 10, pp. 3285–3290, 2012.

[15] F. Iglesias and T. Zseby, "Analysis of network traffic features for anomaly detection," *Mach. Learn.*, vol. 101, nos. 1–3, pp. 59–84, 2015.

[16] C. Callegari, S. Giordano, and M. Pagano, "An information-theoretic method for the detection of anomalies in network traffic," *Comput. Secur.*, vol. 70, pp. 351–365, Sep. 2017.

[17] J. P. Anderson, "Computer security threat monitoring and surveillance," James P. Anderson, Fort Washington, PA, USA, Tech. Rep., 1980.

[18] D. E. Denning, "An intrusion-detection model," *IEEE Trans. Softw. Eng.*, vol. SE-13, no. 2, pp. 222–232, Feb. 1987.

[19] B. Tellenbach, M. Burkhart, D. Schatzmann, D. Gugelmann, and D. Sornette, "Accurate network anomaly classification with generalized entropy metrics," *Comput. Netw.*, vol. 55, no. 15, pp. 3485–3502, 2011.

[20] K. Li, W. Zhou, P. Li, J. Hai, and J. Liu, "Distinguishing DDoS attacks from flash crowds using probability metrics," in *Proc. 3rd Int. Conf. Netw. Syst. Secur. (NSS)*, Oct. 2009, pp. 9–17.

[21] N. Ye and Q. Chen, "An anomaly detection technique based on a chi-square statistic for detecting intrusions into information systems," *Qual. Rel. Eng. Int.*, vol. 17, no. 2, pp. 105–112, 2001.

[22] C. Krügel, T. Toth, and E. Kirda, "Service specific anomaly detection for network intrusion detection," in *Proc. ACM Symp. Appl. Comput.*, 2002, pp. 201–208.

[23] M. Ahmed, A. N. Mahmood, and J. Hu, "A survey of network anomaly detection techniques," *J. Netw. Comput. Appl.*, vol. 60, pp. 19–31, Jan. 2016.

[24] N. Paulauskas and A. F. Bagdonas, "Local outlier factor use for the network flow anomaly detection," *Secur. Commun. Netw.*, vol. 8, no. 18, pp. 4203–4212, 2015.

[25] E. Bigdeli, M. Mohammadi, B. Raahemi, and S. Matwin, "Incremental anomaly detection using two-layer cluster-based structure," *Inf. Sci.*, vol. 429, pp. 315–331, Mar. 2018.

[26] M. Ahmed and A. N. Mahmood, "Novel approach for network traffic pattern analysis using clustering-based collective anomaly detection," *Ann. Data Sci.*, vol. 2, no. 1, pp. 111–130, 2015.

[27] M. F. Umer, M. Sher, and Y. Bi, "Flow-based intrusion detection: Techniques and challenges," *Comput. Secur.*, vol. 70, pp. 238–254, Sep. 2017.

[28] A. Faroughi and R. Javidan, "CANF: Clustering and anomaly detection method using nearest and farthest neighbor," *Future Gener. Comput. Syst.*, vol. 89, pp. 166–177, Dec. 2018.

[29] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, Sep. 1999.

[30] H. Xia, B. Fang, M. Roughan, K. Cho, and P. Tune, "A basisevolution framework for network traffic anomaly detection," *Comput. Netw.*, vol. 135, pp. 15–31, Apr. 2018.

[31] M. L. Proença, Jr., B. B. Zarpelão, and L. de Souza Mendes, "Anomaly detection for network servers using digital signature of network segment," in *Proc. IEEE Adv. Ind. Conf. Telecommun./Service Assurance Partial Intermittent Resour. Conf./E-Learning Telecommun. Workshop (AICT/SAPIR/ELETE)*, Jul. 2005, pp. 290–295.

[32] Y. Guo, X. Chen, and A. Liu, "The research status on meiofauna in China by use of bibliometric analysis," in *Proc. Int. Conf. Challenges Environ. Sci. Comput. Eng.*, Mar. 2010, pp. 507–510.

[33] W. Zhang, J. Gong, W. Ding, and X. Zhang, "Nbos: A fine-grained network management system," *J. Taiyuan Univ. Technol.*, vol. 43, no. 10, pp. 41–46, 2012.

[34] T. Akamatsu, D. Wang, K. Wang, and Y. Naito, "A method for individual identification of echolocation signals in free-ranging finless porpoises carrying data loggers," *J. Acoust. Soc. Amer.*, vol. 108, no. 3, pp. 1353–1356, 2000.

[35] S. Daan and J. Aschoff, "Circadian rhythms of locomotor activity in captive birds and mammals: Their variations with season and latitude," *Oecologia*, vol. 18, no. 4, pp. 269–316, 1975.
[36] M. Dorigo and C. Blum, "Ant colony optimization theory: A survey," *Theor. Comput. Sci.*, vol. 344, nos. 2–3, pp. 243–278, 2005.
[37] M. S. Garshasbi, "Fault localization based on combines active and passive measurements in computer networks by ant colony optimization," *Rel. Eng. & System Safety*, vol. 152, pp. 205–212, 2016.
[38] V. Sharma and A. Grover, "A modified ant colony optimization algorithm (mACO) for energy efficient wireless sensor networks," *Optik-Int. J. Light Electron Opt.*, vol. 127, no. 4, pp. 2169–2172, 2016.
[39] A. Amokrane, R. Langar, R. Boutaba, and G. Pujolle, "Flow-based management for energy efficient campus networks," *IEEE Trans. Netw. Service Manage.*, vol. 12, no. 4, pp. 565–579, Dec. 2015.
[40] E. K. Viegas, A. O. Santin, and L. S. Oliveira, "Toward a reliable anomaly-based intrusion detection in real-world environments," *Comput. Netw.*, vol. 127, pp. 200–216, Nov. 2017.
[41] A. A. Poli and M. C. Cirillo, "On the use of the normalized mean square error in evaluating dispersion model performance," *Atmos. Environ. A, Gen. Topics*, vol. 27, no. 15, pp. 2427–2434, Oct. 1993.

XIAO-DONG ZANG received the M.Sc. degree in computer science and technology from the Nanjing University of Posts and Telecommunications, China, in 2013. He is currently pursuing the Ph.D. degree with the School of Cyber Science and Engineering, Southeast University, Nanjing, China. His research interests include computer networks and security, intrusion detection, network traffic, and host profiling.

JIAN GONG received the B.S. degree in computer software from Nanjing University and the Ph.D. degree in computer science and technology from Southeast University, Nanjing, China, where he is currently a Professor with the School of Cyber Science and Engineering. His research interests include network architecture, network intrusion detection, and network management.

XIAO-YAN HU received the Ph.D. degree from Southeast University, Nanjing, China, in 2015, where she is currently an Assistant Professor with the School of Computer Science and Engineering. Her research interests include future network architecture and network security.