CCF Transactions on Networking
https://doi.org/10.1007/s42045-019-00023-9

REGULAR PAPER

IP backbone traffic behavior characteristic spectrum composing and role mining

Xiaodong Zang[1,2,3] · Jian Gong[1,2,3] · Siyi Huang[1,2,3] · Xiaoyan Hu[1,2,3] · Yun Yang[1,2,3]

Received: 22 May 2018 / Accepted: 29 October 2019

Abstract
The discovery and description of the IP traffic behavior is of great significance for both network operation management and network security monitoring. Researches demonstrate that there are some similarities of the traffic behavior among different IPs, hence, they can be clustered based on the behavior similarity. These similar traffic behaviors can be depicted by a specific behavior pattern called IP address role in our work. Towards this end, a unidirectional IP flow record is used to represent an independent IP activity. The traffic behavior metrics are defined in four dimensions including the duration time, the peer address, the application types and the number of packets and bytes contained in the flow, which corresponds to temporal dimension, spatial dimension, category dimension and intensity dimension, respectively. Nine single-attribute and thirty-nine dual-attribute metrics are extracted from four dimensions to compose the IP address traffic characteristic spectrum, which is used to profile the behavior of all IPs in the observed network and provide the data for the behavior description of each class of IP. These classes are established by a characteristic spectrum matched IP address role mining algorithm designed in this paper. NetFlow data collected from some border routers of China Education Research Network backbone (CERNET) is used to verify the method. Experimental results demonstrate that our approach can be applied to anomaly behavior detection and mainstream behavioral habits analysis.

Keywords IP traffic analysis · NetFlow · Behavior pattern · Characteristic spectrum · Network security

1 Introduction

Network security situation awareness is a cognitive process about the security state of network system. It includes the steps of gradually fusing the original measurement data, extracting the background state and activity semantics of the system and identifying various types of network activities as well as the intention of the anomalous activity, so as to achieve an understanding of the network security situation and the influence of this situation on the normal behavior of the network (Gong et al. 2017). In other words, network security situation awareness requires an overall knowledge about the both normal traffic behavior and abnormal traffic behavior in the network, and a fine-grained traffic behavior description to outline the traffic characteristics with the behavior similarities for all the IP addresses in active, instead of just looking at a specific host or application.

IP traffic behavior analysis is widely applied in anomaly intrusion detection (Umer and Yaxin 2017; Zhao et al. 2013a) and network application classification (Schatzmann et al. 2010; Deri 2012). Anomaly-based IDS refers to finding patterns that are not expected to be normal behavior patterns, while network applications classification based on behavioral characteristics (Chen and Gong 2012), requires the prior understanding of the different behavior characteristics exhibited of the flows or hosts in the interaction with each other. There are two ideas for IP traffic behavior analysis: (1) analyze traffic behavior based on single IP address (Wei et al. 2006; Iliofotou et al. 2010; Karagiannis et al. 2005; Himura et al. 2013) with the aim of revealing the

✉ Xiaodong Zang
xdzang@njnet.edu.cn

Jian Gong
jgong@njnet.edu.cn

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

Published online: 08 November 2019

behavior pattern of single end host; (2) identify a group of end hosts that have similar behavior pattern, which includes two different goals, one is to classify the behaviors in order to comprehensively describe the behavior of all IP addresses (Lee and Brownlee 2007; Xu et al. 2011, 2014; Jakalan et al. 2016a, b; Jiang et al. 2010; Krishna Reddy et al. 2002), the other is attempting to find IPs with specific behaviors (Umer et al. 2017; Kozik 2018; Kheir 2013; Zhao et al. 2013b; Grill et al. 2015; Gañán et al. 2015; Saied et al. 2016; Chen et al. 2007; Hoque et al. 2015; Fayaz and Tobioka 2015), such as DDoS, worm propagation and etc. Our work is closer to behavioral classification among different of IPs. Although, these previous works can summarize the overall profile of IP traffic behavior, they are relatively poor in the realization of fine-grained behavior description, according to the requirement of network security situation awareness.

In an IP network, the traffic behavior of the IP reflects the behavior of the user who uses it in communication. According to the similarity of people's daily activities and social behaviors, the IPs with similar traffic behaviors can be grouped into a cluster. In other words, the IP addresses in same cluster usually show a similar traffic behavior. In order to achieve a more fine-grained IP traffic behavior classification, this paper attempts to set up a more reasonable behavior metrics set to classify the IPs with similar traffic behaviors into clusters. These similar traffic behavior clusters are representative of specific behavior patterns called IP address roles in our paper. A unidirectional flow record of an IP address is considered as an independent IP activity, and the flow traffic behavior metrics in four dimensions are defined, including duration time, the peer addresses, the application types, and the number of packets and bytes contained in the flow, which corresponds to the temporal, spatial, category and intensity dimension, respectively. Nine single-attribute and thirty-nine dual-attribute metrics extracted from four dimensions mentioned above to compose IP address traffic characteristic spectrum to profile the behaviors of all IPs in the observed network. Then, this characteristic spectrum is used to mining different roles with classified the traffic behaviors for all the active IP addresses in the observed network. By interpreting the roles, we find that they contain both normal behaviors and abnormal behaviors. In addition, by applying depth of flow inspection (DFI) on the number of activities the IPs participated contained in the roles on the temporal and spatial dimensions, mainstream behavioral habits of the users can be obtained. The idea proposed in this paper can describe the traffic behavior of IP addresses from both aspects of network security situation perception and network security situation comprehension. 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. The concept of IP address traffic behavior characteristic spectrum is proposed, which is the abstract of all possible values of these metrics, and can be used to profile the behavior of all IPs in the observed network. The characteristic spectrum construction algorithm is given.
3. Based on the characteristic spectrum, the concept of IP address role is proposed, which represents a specific traffic behavior pattern of a group of IP addresses. An efficient characteristic spectrum matched IP address role mining algorithm is given.
4. Experiment results with actual measured dataset show that the IP address roles obtained in this paper can be applied to abnormal behavior detection and mainstream behavioral habits analysis.

The remainder of the paper is organized as follows. Section 2 introduces related algorithms of traffic analysis. Section 3 describes IP address traffic behavior metrics set. Section 4 describes the algorithm to construct IP address traffic behavior characteristic spectrum. Section 5 provides a simple technique to mining and interpret the IP address roles of observed network. In Sect. 6, we validate our approach by applying the described techniques on CERNET backbone data. The conclusion of the paper is given in Sect. 7.

2 Related work

With the continuous growth in the scale and diversity of Internet hosts and applications, it is becoming increasingly important to understand traffic patterns of end-hosts and network. Discovering and describing IP address traffic behavior have a great significance for network operation management and network security. Moreover, researches demonstrate that there are some similarities exist in the traffic behavior of different IP addresses, which has attracted significant attention of network researchers. However, this issue is far from being solved.

There are currently two ideas for IP addresses traffic behavior analysis. On one hand, analyze traffic behavior based on a single IP address (Wei et al. 2006; Iliofotou et al. 2010; Karagiannis et al. 2005; Himura et al. 2013) with the goal to reveal the behavior pattern of single end-host. Illiofotou et al. (2010) exploited IP communication graph and the information of some applications with the purpose of profiling Internet backbone traffic, while Karagiannis et al. (2005) studied the host behavior at the social, functional and application levels to classify traffic flows. On the other hand, identify a group of different hosts that have similar behavior pattern. This idea includes two different goals, one is to classify the behaviors in order to fully describe the behavior of

all IP addresses (Lee and Brownlee 2007; Xu et al. 2011, 2014; Jakalan et al. 2016a, b; Jiang et al. 2010; Krishna Reddy et al. 2002), such as (Lee and Brownlee 2007) merged packet data into clusters based on the similarity of network traffic characteristics, but it was unavailable when the packets were encrypted or from aggregated flow records. However, (Xu et al. 2011; Jakalan et al. 2016a, b) proposed an approximation algorithm to discover communities on a large scale in the managed domain based on the bipartite networks and one mode projection graph. The difference is that Xu et al. (2011) explored the behavior similarity of Internet end hosts in the same network prefixes, while Jakalan et al. (2016a, b) proposed method to explicitly address location information (inside/outside) for IP clustering by splitting the entire IP address space into inside (the managed domain) and outside ones and the clustering method is used to discover groups of inside IP addresses that communicate with common outside IP addresses. In addition, Jiang et al. (2010) described structural patterns in the overwhelming network using the characteristics including daily aggregate traffic volume, traffic distribution in time and space, traffic distribution in application and traffic balance in flow direction etc. The other one is trying to find IP addresses with specific behaviors (Umer et al. 2017; Kozik 2018; Kheir 2013; Zhao et al. 2013b; Grill et al. 2015; Gañán et al. 2015; Saied et al. 2016; Chen et al. 2007; Hoque et al. 2015; Fayaz and Tobioka 2015). Kheir et al. (2013) analyzed anomalies in the HTTP user agent with the aim to reveal valuable detection patterns that used to classify user agent anomalies and to extract signatures for malware detection. Chen et al. (2007) developed a distributed change-point detection architecture by using change aggregation trees to detect DDoS attack across multiple network domains at the earliest time in order to minimizes the flooding damages to the victim systems.

Our work is closer to behavioral classification among different of IPs. Although, behavioral classification of different IPs can summarize the overall profile of the IP traffic, it is relatively poor in the realization of fine-grained behavior description, e.g., analyzing the behavior pattern of a specific application or traffic behavior is always active in a certain period of time, etc. In addition, we find that the measures are not comprehensive, resulting in incomplete behavioral analysis. For instance, Jiang et al. (2010) extracted the volume of traffic in different attributes as a measure (flow size, ratio of upload and download packets in streams, and stream size distribution in different applications, etc.), which only describes traffic behavior from the perspective of traffic intensity, however, (Xu et al. 2011; Jakalan et al. 2016a, b) analyzed the communication relationship of the IP address to describe the traffic behavior of the IP address from the spatial point of view, without considering the temporal relationship of the traffic behavior.

This work explores the similarity behavior of IP address from four dimensions including temporal, spatial, category and intensity. To the best of our knowledge, this work gives a more comprehensive metrics set to characterize IP traffic behavior for the first time. Additionally, IP address traffic characteristic spectrum is first proposed to profile the behavior of all IPs in the observed network, and to provide the data for the behavior description of each IP. The derived IP address roles can be applied to the applications of anomaly behavior detection and mainstream behavioral habit analysis.

3 Extracting IP address traffic behavior metrics

In order to systematically and comprehensively describe the traffic behavior of the IPs, forty-eight metrics are extracted, including nine single-attribute and thirty-nine dual-attribute metrics from four dimensions. Single-attribute metric refers to the metric involving only one dimension. Though, 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. Dual-attribute metric refers that there are two dimensions exist in a metric, which is divided based on another dimension, such as *DvsInScp* divided based on spatial dimension represents the number of application types contained in the flows. The calculation methods and the semantics of these metrics are described below.

Table 1 Single-attribute metrics and their semantics

Dimensions	Metrics	Semantics	Calculate formulas
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)
Intensity	*Ocr*	The numbers of flows	Formula (5)
	IstSendPkts	The number of packets send of the flow	Formula (6)
	IstSendByts	The number of bytes send 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)

3.1 Extracting single-attribute metrics from IP flows

NetFlow is a traffic monitoring tools in the Internet nowadays, 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 IP flows used in this paper are collected from the border of CERNET in Jiangsu Province, including attributes of 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 application type (*Toa*), the number of packets (*Pkts*), the number of bytes (*Bytes*), the number of packets send from source to destination IP (*l_pkts*), the number of bytes send from source to destination IP (*l_bytes*), the number of packets send from destination to source IP (*r_pkts*), the number of bytes send 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 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 refer to the association between the activity and the location, and use the degree of association to express. In an IP network, the *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 (Cui et al. 2017), denoted as activity diversity (*Dvs*). In an IP network, *Dvs* is equal to the number of application types (Li et al. 2013) that the IP flow record belongs to. The application type of an IP flow record of this article originates from the Network Behavior Observation System (NBOS). NBOS is a network traffic behavior monitoring system for monitoring and managing CERNET's service quality and security status (Weijie 2010). The source and destination ports, upper layer protocol, packet arrival interval and bidirectional packet ratio, etc., are incorporated to identify the application types of the IP flow record. Assuming that P is the number of flows in an observed cycle.

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

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*). In an IP network, the more packets and bytes send or received, the greater influence on the network performance. There are four metrics used to describe the activity intensity of the IPs, such as the number of packets send (*IstSendPkts*), the number of bytes send (*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 send from the observed IP to the peer IP address.

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

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

In summary, the semantics and calculation method of all single-attribute metrics are shown in Table 1.

3.2 Extracting dual-attribute metrics from IP flows

This article refers the analogous method of animal ecology when constructing dual-attribute metrics of IP traffic behavior. 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 (Zheng 2005). Similarly, the IP activities are grouped into N subsets in terms of the temporal attribute, 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 respectively.

3.2.1 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 falls into the unit time is divided into a subset and use it to describe the changes of other divided attributes. By using unit time based 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 the activity range, and is used to describe whether the activity range is changing rapidly or not. (2) Dividing by period of time,

which is used to characterize the rhythm behavior of the IP. The rhythm behavior of an IP address refers to the periodicity of the users' behavior reflected by IP address traffic, e.g., whether it is active during the daytime or at night, or whether the behavior often appears in a specific period of time, and etc. In this paper, 1 day is divided into six periods of time to discovery their rhythm behaviors. Metrics in each period of time are counted, and their semantics as shown in Table 2.

3.2.2 Dual-attribute metrics division based on spatial

Two division methods are also considered, on the one hand, dividing the flow based on peer IP address. This dividing approach can be used to depict the degree of association of the observed IP, which correspond to the specific location it communicates with. On the other hand, 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 observed IP, the wider the communication range is. The dual-attribute metrics division based on spatial dimension and their semantics are given in Table 3.

3.2.3 Dual-attribute metrics division based on category

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

4 Composing IP address traffic behavior characteristic spectrum

In order to profile the behaviors of all IPs in the observed network and provide data for the behavior description of each IP, the concept of IP address traffic behavior characteristic and its calculation method are given first. Then, the characteristics are discretized by using ID3 algorithm. Finally, the algorithm of constructing and updating IP address traffic behavior characteristic spectrum are given.

4.1 Calculating IP address traffic behavior characteristics

Definition 1 IP address traffic behavior characteristic refers to the abstraction of the different behavioral traits of the IP address during the communication process, which is specifically represented by the form of the value of 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

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

Divided method	Dimension	Metrics	Semantics	Calculate formulas
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)
	Intensity	*OcrPerUnit*	The number of flows per unit time	(5)
		IspPerUnit	The number of packets send per unit time	(6)
		IsbPerUnit	The number of bytes send per unit time	Ref. (6)
		IrpPerUnit	The number of packets received per unit time	Ref. (6)
		IrbPerUnit	The number of bytes received per unit time	Ref. (6)
Divided by time periods	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)
	Intensity	*OcrInRtm*	The number of flows in each period of time	(5)
		IspInRtm	The number of packets send in each period of time	(6)
		IsbInRtm	The number of bytes send 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

Divided method	Dimension	Metrics	Semantics	Calculate formulas
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)
	Intensity	*OcrPerNdg*	The number of flows of the observed IP communicate with each peer IP	(5)
		IspPerNdg	The number of packets send between the observed IP and the peer IP	(6)
		IsbPerNdg	The number of bytes send 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)
	Intensity	*OcrInScp*	The number of flows of the observed IP communicate with each peer IP segment	(5)
		IspInScp	The number of packets send between the observed IP and each peer IP segment	(6)
		IsbInScp	The number of bytes send 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)

Table 4 Dual-attribute metrics division based on category dimension and their semantics

Divided method	Attributes	Metrics	Semantics	Calculate formulas
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)
	Intensity	*OcrPerClss*	The number of flows of each type of flow	(5)
		IspPerClss	The number of packets send of each type of flow	(6)
		IsbPerClss	The number of bytes send of each type of flow	Ref. (6)
		IrpPerClss	The number of packets received of each type of flow	Ref. (6)
		IrbPerClss	The number of bytes received of each type of flow	Ref. (6)

is the metric value. ② Dimension-determined vector, a set of values are used to describe the characteristics of the IP traffic behavior. ③ Size-uncertain set, the statistical values including the average, maximum and summation 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.

4.1.1 Calculating the behavior characteristics for single-attribute metrics

Two kind of numerical value forms 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*, *Scp* and etc. with the numerical value of the each metric as their characteristic value. While, size-uncertain set are calculated multiple times throughout the observed cycle with the same meaning of each calculation, such as *IstSendPkts* and *IstSendByts*. Then, the average, maximum and summation are used to summarize the meaning of the metrics, in the meantime, uses them as their characteristic value.

4.1.2 Calculating the behavior characteristics for dual-attribute metrics

The numerical value form of each dual-attribute metric is affected by different division method. The approach of dividing all flows into a size-fixed flow subsets according to rhythmicity and flow application types is used. In order to describe the rhythmicity behavior of each IP, 1 day is divided into six periods of time, and put the flows with their start time fallen in each period 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 numerical characteristics of each position vector in the set, such as *IspInRtm*, *IrpInRtm*, *IspPerClss*, *IsbPerClss*, *IsbInRtm*, *IrbInRtm*, *IrpPerClss* and *IrbPerClss*.

4.2 Discretizing IPs' traffic behavior characteristics

After obtaining the IP addresses' traffic behavior characteristics, we find that they are randomness, large volume and noise. Using such data directly for data mining or learning not only consumes large amount of resources, but also overwhelms the effective or useful information in the data. In order to solve this problem, the numerical characteristics are discretized. The process of characteristic discretization is to select appropriate breakpoints to spilt the characteristic range, and convert the numerical values that fall within each range to specific symbols with the aim of achieving data reduction. In this paper, we use the ID3 algorithm based on information entropy to discretize the traffic behavior characteristics. In the process of discretization, our goal tends to assign similar features of the IP address to the same interval without involving the classification of IP addresses.

The experimental data of this paper originated from "67,584" different IPs, therefore, the number of characteristics are too large, which result in a large number of symbol features and cannot achieve the goal of characteristics reduction. Besides, the key of the algorithm is to select the appropriate breakpoint to separate the samples into different interval as much as possible. For these reasons, in the process of discretization, we use the minimum occurrence of characteristics as the threshold to avoid overly detailed division of characteristic values. Each metric is discretized separately by using the pseudo code in algorithm 1. The input "$(f_1, f_2, \ldots, f_n)$" represent all possible behavior characteristic values of each metric. The output "$C_1(f_1 - f_i), \ldots, C_j(f_j - f_{j+k})$" are the set divided by the breakpoint, each of which contain a list of characteristic values.

Algorithm 1: The pseudo code of discretizing IP address traffic behavior characteristics

Input: IP addresses' traffic behavior characteristic values $(f_1, f_2, \ldots, f_n)$ in an IP network

Output: $C_1(f_1 - f_i), \ldots, C_j(f_j - f_{j+k})$

① With root as the pending node T, put all the characteristic values into root.

② Sort the characteristic values in T in ascending order.

③ Initialize the set of breakpoints by taking the mean of neighboring numerical features as a breakpoint.

④ Perform breakpoint screening, for each dividing interval, if the total number of occurrences of the characteristic value in the interval is greater than the threshold, use it as a candidate breakpoint, otherwise delete it.

⑤ If the candidate set of breakpoints is not empty, the information gain of each breakpoint is calculated and the breakpoint with the largest information gain is selected to divide the characteristics value domain on the node, as a subtree of T, respectively.

⑥ Return to step ② to handle the child node T.

Taking breakpoint "A" as an example, its information gain is calculated as follows:

Put the behavior samples associated with all the characteristic values in the node T to be processed into the set S, the size of which is denoted as |S|. In the observed IPs, the behavior sample sets for each IP address are C_1, C_2,…, C_n, the information entropy of the sample set S is:

$$\mathrm{E(S)} = -\sum_{i=1}^{n} p(Si) \times \log_2(1/p(Si)) \tag{7}$$

$p(Si) = |C_i|/|S|$, indicates the probability of the sample of the IP address i. Assume that the breakpoint A divides S into two different sets {S1, S2}. The conditional entropy of A is calculated as follows:

$$\mathrm{E(S|A)} = \sum_{j=1}^{2} P(Sj) \times \sum_{i=1}^{n} P(Ci/Si) \times \log_2(1/p(Ci/Si)) \tag{8}$$

$p(Ci|Sj)$ is the probability of Ci. If A can reduce the randomness of behavior samples belonging to different IP addresses, then, E(S|A) < E(S).

$$\mathrm{Gain(A)} = \mathrm{E(S)} - \mathrm{E(S|A)} \tag{9}$$

4.3 Composing IP address traffic behavior characteristic spectrum

In animal ethology, the ethogram 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 (Xiao and Wang 2005) are required to establish it. Not only can it be used to describe the behavior rules of all animals, but it can also be applied to analyze individual behaviors. According to the concept of ethogram, the definition of the IP address traffic behavior characteristic spectrum is given as follows.

Definition 2 IP address traffic behavior characteristic spectrum is the full record of the stable characteristics obtained from the IPs 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.

There are two functions exist of the IP address traffic behavior characteristic spectrum. (1) IP address traffic behavior characteristic spectrum is an actual observation of the traffic behavior in the managed network, which can profile the behavior of all IPs in the observed network; (2) different IPs in the network have different characteristics, therefore, the behavior of a single IP address can be analyzed through role mining, as discussed in the following section. The stability of IP address traffic behavior characteristics is divided into existence stability and association stability. The existence stability of the behavior characteristic refers to the long-term nature of the managed network, regardless of the characteristics occur on the specific IP. 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 frequency of the occurrence of the characteristic is used as the stability factor. The association stability of the behavior reflects whether the characteristic is associated to the IP for a long time or not, and use association rate to measure it. Association rate is equal to the ratio of the associated duration time to the total active length of

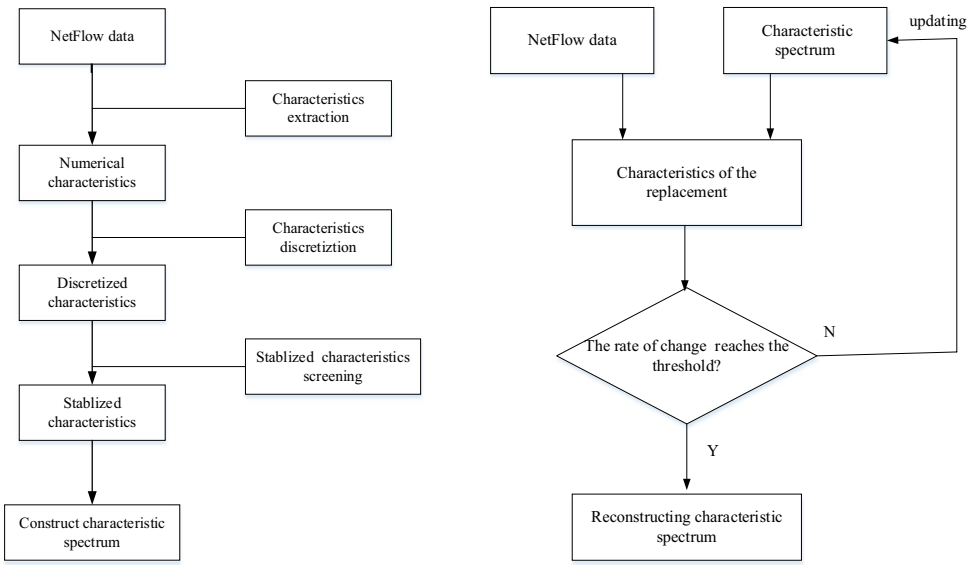

Fig. 1 The process of composing (left) and updating (right) IP traffic behavior characteristic spectrum

IP address. A low association rate indicates that the stability between the characteristic and the IP address is not strong. The characteristics with low frequency and low association rate are eliminated when constructing the characteristic spectrum. The steps of constructing IP address behavior characteristic spectrum in one observed cycle include characteristic value extraction, numerical value characteristic discretization and stability characteristics screening, as shown in Fig. 1 (left).

As the applications and users in the network change 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 (Fig. 1 right). According to the degree of network traffic changes, the updating process including the following operations:

① Replacement of characteristics. When the network traffic is slightly fluctuating, or a few IPs' traffic behaviors have changed, in this case, some characteristic in the characteristic spectrum are no longer stable. Hence, only the stability of discretized characteristics needs to be calculated, and replaces the characteristics that are not stable in the characteristic spectrum.

② Reconstruction of the characteristic spectrum. When the network structure has changed, or the DDoS attack appears, in this case, a large number 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 are not applicable and the reconstruction of the characteristic spectrum is required.

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 characteristic account for the ratio in the current characteristic spectrum called the rate of change. When the rate of change is sufficiently large, in this case, the characteristic maintenance rate is small, which indicates most of the IP traffic behaviors have changed, and it is necessary to reconstruct the characteristic spectrum.

5 Mining and interpreting IP address traffic behavior roles

The concept of IP address traffic behavior role and the problems faced in its construction are given first. Then, IP address role mining algorithm based characteristic spectrum matched is proposed. Finally, we interpret the roles obtained and discuss their applications in the paper.

5.1 IP address roles based on characteristic spectrum

Definition 3 IP address traffic behavior role refers to similar behavior pattern reflected by multiple IP addresses during their communications in the managed network.

Two problems are considered when constructing IP traffic behavior role, including the storage form and which characteristics should be stored. As for the former one, since the original characteristics are in an over-detailed numerical form, misjudgment of different characteristics can be caused due to merely subtle fluctuations of the metric values, which is not conducive for roles mining. As for the latter, due to the randomness nature of users and the network itself, some characteristics may no longer occur after one appearance. In other words, these characteristics can't provide accurate semantic information. These problems can be solved by composing IP traffic behavior characteristic spectrum, as it records all the stable traffic behavior characteristics. IP address role mining method based characteristic spectrum matched is designed to discover individual IP behavior.

According to the analysis of the stability of the characteristics in the previous section, the association stability between the IP and the characteristic is different, and measured by its association rate. If the association rate is equal to one between the characteristics and the IP within a single observed cycle, which indicates that during this period, this characteristic must be appeared in the IP traffic behavior. Otherwise, if the association rate is low, which indicates that this characteristic only appear with a certain probability during this period of time. A fixed characteristic subset and the accidental characteristic subset are designed to represent the structure of IP address role.

Characteristics in a fixed characteristic subset associated to the role's IP address with the association rate is one during a certain observed cycle, i.e., IP address i will reveal all elements within role B's fixed characteristic subset if it belongs to role B. On the contrary, characteristics of an accidental characteristic subset associated to the role's IP address with a relatively low rate. Elements within the accidental characteristic subset of role C can only be observed with a certain probability from IP address j, if j belongs to

role C. The meaning of IP address belonging to a specific role can be interpreted as the stable characteristics associated to it must be found in the role's fixed characteristic subset, while the remaining characteristics may be found within the role's accidental characteristic subset with only a certain probability.

5.2 IP address role discovery using sliding window

Although, the traffic behavior of IPs is not always static in a short period of time, their behaviors can maintain a fixed behavior pattern. However, the access of new users, the changed network structure and the occurrence of DDoS attacks would result in a great change of their behaviors. Besides, as time passes, new flows continue to emerge,

Table 5 IP address role A and B's feature table

Attributes	Metrics and semantics	Role A's characteristic subset		Role B's characteristic subset	
		Fixed	Accidental	Fixed	Accidental
Temporal	*Drt*				
	Total duration time(s)	26,189–17,949,600			117–4518,4519–7603, 7604–13,023
	Average duration time(s)	58–167			29–57, 58–167
	Max duration time(s)	299			75–285, 285 –299
Spatial	*Ndg*	1050–8210			16–130, 31–1049
	Scp	828–3745			13–120, 21–827
Category	*Dvs*	1–5		1–5	
Intensity	*Ocr*	23,172–119,709			3–30,31–250,251–1085
	IstSendPkts				
	Total number packets send	25,856–6.79×10^7			768–2816, 3072–8192
	Average number packets send per flow		158–703, 704–5235, 5236–122,252	1–20	
	Max number packets send per flow	25,120–3.49×10^6		256–512	
	IstSendByts				
	Total number bytes send	2.35×10^6–4.87×10^9			1.53×10^4–4.65×10^4, 4.68×10^4–2.93×10^5, 2.94×10^5–7.71×10^5
	Average number bytes send per flow	20,970–3.09×10^7			1–1534, 1536–5386
	Max number bytes send per flow	462,336–1.27×10^9			1.53×10^4–1.99×10^4, 2.02×10^4–5.99×10^4, 6.04×10^4–2.25×10^5
	IstReceivePkts				
	Total number packets received	94,976–1.26×10^8			1024–1.04×10^4, 1.07×10^4–2.04×10^4, 2.07×10^4–9.47×10^4
	Average number packets received per flow		277–339, 340–418, 419–661	219–356	
	Max number packets received per flow	9728–3.41×10^6			7.01×10^4–3.67×10^6, 3.67×10^6–1.72×10^7
	IstReceiveByts				
	Total number bytes received	1.23×10^8–1.89×10^{11}			219–276,276–349
	Average number bytes received per flow		1.39×10^4–1.30×10^5, 1.30×10^5–3.13,10^5, 4.41×10^5–7.96×10^5	1.39×10^4–1.30×10^5	
	Max number bytes received per flow	1.28×10^8–4.85×10^9		1.48×10^4–1.14×10^5	

therefore, a stream processing method is required for IP roles mining and updating.

To avoid processing an enormous volume of traffic data, a single-cycle feature set is built to store the characteristics that appear on the IP. In this paper, a size-fixed time window is adopted. When the observed window is fully filled, the IP role discovery can be performed. Taking IP address *i* as an example, configure n-days as window size with daily observed once. From the first day of finding *i*, collecting the characteristics of *i* that belong to the characteristic spectrum to establish a single-period feature set. After n-days observation, performing role discovery. Finally, updating the flow records in the window, and checking the validity of current role *i*. If the role is invalid, updating the role with the latest IP flow records. The specific method of role discovery is described as follows.

① For each IP, traversing all the single-period feature sets where the IP locates first, and collecting all the occurred features. Then, placing them into the feature sets that the role of the IP belongs to, and recording the occurrence times of each feature.
② Calculating the association rate of features in the feature set. Putting the features with the association rate of one in the fixed feature subset and placing the other features with low association rate in the accidental feature subset to construct the behavior role of the IPs. Association rate is equal to the number of feature appearance times to the size of time window. In this paper, "67,584" IPs on 10 network segments of 3 units in Jiangsu Education and Research Network is employed. It has obtained 130 roles for five consecutive days. Table 5 shows the 202.195.***.*** and 202.195. ***.*** corresponding to the roles A and B, where B contains twelve IPs and A only one IP.

It is easy to see that role B has fewer fixed features and most features stored in accidental feature subset. Comparing with role A, role B has shorter active time and smaller scope of communication, which has less influence.

5.3 Interpreting IP address roles

After analyzing the obtained IP address roles, we find that although it can describe the traffic behavior of IPs, it lacks intuitiveness and relativity. Intuitiveness refers to when querying the IP roles, they can help administrators understand the behavior characteristics in both fixed and accidental characteristic group quickly. Relativity means that although the role records the numerical range of each metric, it is not

Table 6 The numerical range and symbols of the characteristic spectrum

Metrics	Means of obtaining characteristics values	Range of numerical characteristic	Ranges
Drt	Sum	117–4518, 4519–7603, 7604–13,023, 13,025–26,186, 26,189–17,949,600	5
	Avg	29–57, 58–167, 168–276	3
	Max	0–75,75–284, 285–298, 299	4
Ndg	Value	3–15, 16–130, 131–1049, 1050–8210	4
Scp	Value	2–12, 13–120, 121–827, 828–3745	4
Dvs	Value	1–5, 6–7, 8–10, 11–14	4
Ocr	Value	3–30, 31–250, 251–1085, 1086–23,171, 23,172–119,709	5
IstSendPkts	Sum	256–512, 768–2816, 3072–8192, 8448–25,600, 25,856–67,965,696	5
	Avg	1–20, 21–157, 158–703, 704–5235, 5236–122,252	5
	Max	256–512, 768–6536, 6792–24,864, 25,120–3,495,420	4
IstSendByts	Sum	15,360–46,592, 46,848–293,888, 294,400–771,072, 771,584–2,355,200, 2,356,220–4,878,080,000	5
	Avg	1–1534, 1536–5386, 5389–9754, 9757–20,969, 20,970–30,930,200	5
	Max	15,360–19,968, 20,224–59,904, 60,416–225,280, 225,536–461,824, 462,336–1,274,099,968	5
IstReceivePkts	Sum	1024–10,496, 10,752–20,480, 20,736–39,680, 39,936–94,720, 94,976–126,754,000	5
	Avg	219–276, 277–339, 340–418, 419–661, 662–125,357	5
	Max	256–512, 768–1280, 1536–3328, 3584–9472, 9728–3,409,150	5
IstReceiveByts	Sum	70,144–3,672,060, 3,672,580–17,216,500, 17,245,700–38,770,200, 38,776,100–102,603,000, 102,686,000–188,794,994,688	5
	Avg	13,952–130,247, 130,283–313,400, 313,403–441,266, 441,280–796,983, 797,040–177,759,008	5
	Max	14,848–1,147,390, 1,148,670–1,922,560, 1,926,140–4,613,890, 4,614,140–13,788,200, 13,842,400–4,851,850,240	5

Table 7 IP address role interpretation table

Attributes	Metrics and sematics	Role A	Role B
Temporal	*Drt*		
	Total duration time (s)	5/5	1/5, 2/5, 3/5
	Average duration time (s)	2/3	1/3, 2/3
	Max duration time (s)	4/4	3/4, 4/4
Spatial	*Ndg*	4/4	2/4, 3/4
	Scp	4/4	2/4, 3/4
Category	*Dvs*	1/4	1/4
Intensity	*Ocr*	5/5	1/5, 2/5, 3/5
	IstSendPkts		
	Total number packets send	5/5	1/5, 2/5
	Average number packets send per flow	3/5, 4/5, 5/5	1/5
	Max number packets send per flow	4/4	1/4
	IstSendByts		
	Total number bytes send	5/5	1/5, 2/5, 3/5
	Average number bytes send per flow	5/5	1/5, 2/5
	Max number bytes send per flow	5/5	1/5, 2/5, 3/5,
	IstReceivePkts		
	Total number packets received	5/5	1/5, 2/5, 3/5
	Average number packets received per flow	2/5, 3/5, 5/5	1/5
	Max number packets received per flow	5/5	1/5, 2/5
	IstReceiveByts		
	Total number bytes received	5/5	1/5, 2/5
	Average number bytes received per flow	1/5, 2/5, 4/5	1/5
	Max number bytes received per flow	5/5	1/5

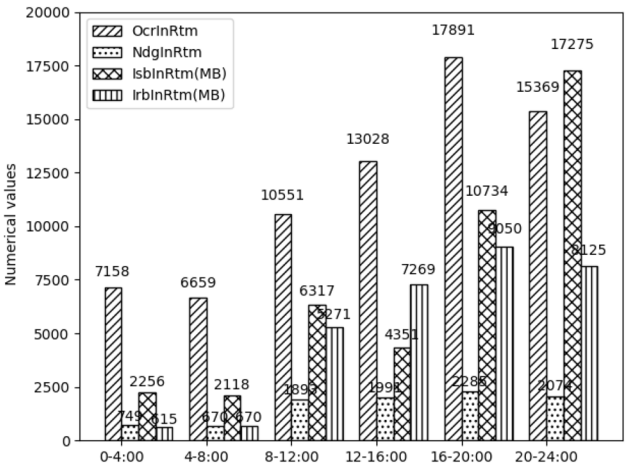

Fig. 2 The traffic behavioral characteristic spectrum

easy to understand their meanings. Therefore, the interpretation method of the IP address role is given, then, the applications of IP role is discussed, such as constructing the profile of IP network, detecting anomalous behavior and analyzing mainstream behavioral habits of users.

According to the Definition 3, the interpretation of IP address roles is essentially the interpretation of their characteristics, which are derived from 48 metrics. Each metric can describe a specific property of IP, e.g., *Drt* is used to describe the duration time of IP traffic behavior. The larger the numerical value is, the stronger the corresponding property is. Table 6 gives the numerical value range distribution of the discretized characteristics for facilitate querying. Two discretized algorithm are used, experiment result shows that ID3 algorithm is better than the unsupervised equal frequency binning (FRQ) algorithm. This paper illustrates the difference of the traffic behavior by interpreting and comparing the IP roles according to numerical value distribution, for instance, the range of *Dvs* is 1–15, and is divided into four ranges further, namely "1–5", "6–7", "8–10" and "11–14" to describe differences of application types. The range of "1–5"is denoted as "1/4", and others range analogously as shown in Table 7. For each row in the table, the features in the role will labeled as fixed if it only has one symbolized characteristic, such as 5/5 of the *Drt* belongs to role A in Table 7, otherwise it is treated as an accidental feature.

5.3.1 Constructing the profile of IP network

In order to profile the behavior of all IPs in the observed network, the flow data of a network segment with address

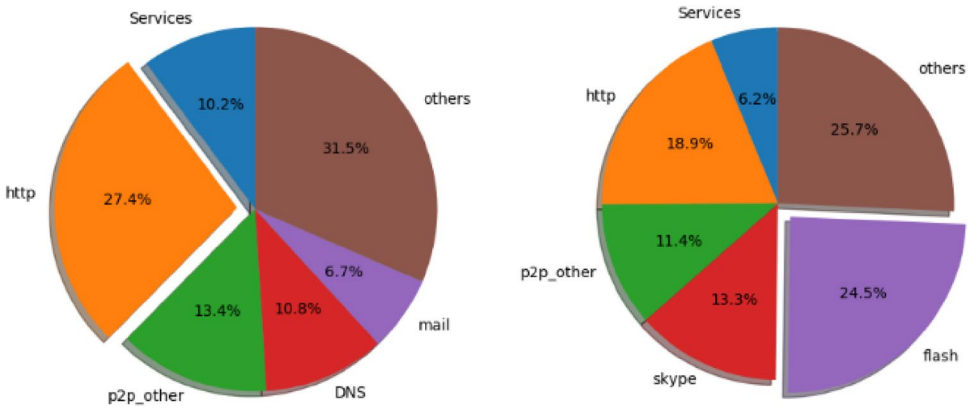

Fig. 3 The traffic composition of different applications, 8:00–12:00 o'clock (left), 16:00–22:00 o'clock (right)

prefix 202.194.*.*/20 from China Education and Research Network is analyzed. Three cycles (one cycle per week) are observed continuously to construct IP address characteristic spectrum of the network segment. Due to space limitations, we do not give the complete characteristic spectrum of the single-attribute metrics and the dual-attribute metrics for the network segment, and only uses the temporal division method to analyze the behavioral changes of the network segment. Six periods of time are separated in 1 day, after continuous observation of three cycles, we found that the traffic behavior of the network segment is basically stable with smaller fluctuations. Average value is calculated among the statistics of the three cycles as the final results, which shows in the Fig. 2.

It can be observed that the network behavior characteristics reflects the user's behavior habits in the network. The volume of the traffic is large and the IP communication range is wider in the network between the afternoon and midnight from Fig. 2. However, the volume of the traffic and IP communication range 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 some absentees contribute to a relatively small amount of traffic. However, in the evening, online learning or entertainment become more, and the traffic is relatively huge. In order to further verify our analysis, we calculate the percentage of specific applications of the flows between 8:00–12:00 and 16:00–22:00. After tracing the applications of the flow in NBOS system, which has a total of 16 application types, 15 different application types for this period of time have been found. The traffic behavioral characteristics account for each different application are shown in Fig. 3. From the traffic statistics chart, we can see the traffic of http, p2p_other, DNS, etc. are relatively large at 8:00–12:00 (Fig. 3, left). It is inferred that users browse news, blogs, mails, and other activities during this period of time, which is consistent with the behaviors of the staff in the campus. However, the traffic of http, flash, skype application at 16:00–22:00 is relatively large (Fig. 3, right), which infers that users in this period are engaged in web services, video, social activities and other entertainment activities that are closer to the students' behavior.

5.3.2 Detecting anomalous behavior

In the context of network, for the purpose of ensuring the integrity, availability and resilience of the network infrastructures, it is necessary not only to know the usage of the entire network, but also to identify and analyze the traffic behavior of IP hosts, which has a great significance for the analysis of hotspot traffic and abnormal behavior (Miao et al. 2015; Marnerides et al. 2014). In Table 7, the differences are found in temporal, spatial, category and intensity dimensions between role A and role B. In temporal dimension, role B is less persistent than role A and the duration time of the corresponding IP activity is relatively weak. The number of IPs is not fixed, but the range is relatively wide. Their traffic intensity reveals that both of their sent and received traffic is relatively small. Besides, ten out twelve of the observed IP correspond to a communication port of 445, which is known to be vulnerable. According to these analysis, the IP's behavior in the role B conforms to horizontal scanning behavior. On the other hand, role A has a fixed strongest and maximum persistence, that is, the corresponding IP behavior

Table 8 Mainstream unit time activity frequency characteristic group

Proportion of associating IPs	Metrics and sematic		Discretized range
21.5%	Number of activities per unit time	The average number of activities per unit time	1/5
		The maximum number of activities per unit time	1/5, 2/5, 3/5, 4/5

of the role in the observed period is always active. In spatial dimension, role A has the strongest association and the widest communication range. In the category dimension, the diversity of the activity of role A has been kept at the lowest level. The types of traffic activities involved are also more concentrated. Statistics demonstrate that 84.8% of the traffic is http. Overall, the IP in role A is active in the network for a long time. The influence is relatively large and extensive, its flow activity has a greater regularity, its application types in the network are narrow and most of them are http applications, which implies that it is more conform to the characteristics of the webserver.

5.3.3 Analyzing mainstream behavioral habits of users

Although the characteristics in the IP address role can clearly describe the users' Internet surfing habits, it is difficult to find two users with approximately same habits in the managed network. In other words, there is no completely identical IP roles. However, for one single dimension, such as temporal and spatial, there may exist some users with similar behavioral habits. Therefore, administrators can describe user behavior habits by intercepting the characteristics as needed. The number of activities in which the IPs participate at temporal and spatial dimensions among the IP roles are analyzed in this part. The characteristic groups with the most associated IPs are considered as the mainstream feature group to describe the mainstream behavioral of the users.

Number of activities feature group of IP per unit time Thirty different feature groups are obtained after dividing the IP roles by the number of activities per unit time. The proportion of the IP addresses associated with each feature group is counted and ranked after numbering. The feature group with the largest area associated with the most IPs accounting for 21.5%, called the mainstream number of activities feature group of IP per unit time. Then, the fixed features and accidental features corresponding to this group are listed, as shown in Table 8. From the table, it can be found that the IPs in this group with very few average number of activities per unit time. In addition, the maximum

Table 9 Mainstream rhythmic activity characteristic group

Proportion of associating IPs	Metrics and sematic	Discretized range
	The number of rhythm activities	
15.8%	Number of activities between zero and four o'clock	2/4, 3/4
	Number of activities between four and eight o'clock	1/4, 2/4
	Number of activities between eight and twelve o'clock	1/4
	Number of activities between twelve and sixteen o'clock	1/4,
	Number of activities between sixteen and twenty o'clock	2/4, 3/4
	Number of activities between twenty and zero o'clock	4/4

Table 10 Mainstream with single address communication activity characteristic group

Proportion of associating IPs	Metrics and sematic		Discretized range
17.9%	Number of communications with a single IP	Average number of flows communicating with a single IP	1/5
		Maximum number of flows communicating with a single IP	1/5, 2/5, 3/5
	Rhythmic communication times with a single IP	Number of communications with a single IP at 0–4 o'clock	1/5
		Number of communications with a single IP at 4–8 o'clock	1/5
		Number of communications with a single IP at 8–12 o'clock	1/5, 2/5
		Number of communications with a single IP at 12–16 o'clock	1/5, 2/5
		Number of communications with a single IP at 16–20 o'clock	1/5, 2/5, 3/5
		Number of communications with a single IP at 20–24 o'clock	4/5
	Communication strength with a single IP	Send packets strength with a single IP	1/4, 2/4, 3/4
		Received packets strength with a single IP	1/4, 2/4, 3/4

number of activities fluctuate between low and large. This result shows that such users are subtler and participate in more activities in a short time, but are not dense.

Rhythmic activity times feature group Forty-five different feature groups are obtained according to the number of rhythmic activities. After numbering, the proportion of the IP addresses associated with each feature group is counted and ranked. The feature group with the largest area is associated with the most IPs accounted for 15.8%, called the mainstream rhythm activity times feature group. Then, the fixed features and accidental features corresponding to the feature group are list in Table 9. We can see that the IP address associated with this feature group participates in activities is less during the day, but more active at night, which has obvious characteristics of being nocturnal.

Communication with a single IP feature group A total of twenty-seven different feature groups are obtained after mining the features in IP address roles that communicate with a single IP. The proportion of the IPs associated with each feature group is counted and ranked. The corresponding fixed features and accidental features in the feature group with the largest proportion (17.9%) are listed in Table 10. Although, having analyzed the communication times of IPs in this feature group with a single IP, we do not obtain valuable information. In order to further study its behavior, we do the statistics of the rhythm behavior of IPs communications with a single IP and the communication intensity in this feature group. The finds are that these IPs have more communication time at 20:00–24:00 and the traffic intensity is variable but not strong. We consider that these IPs should be engaged in point-to-point chats or short-term video behaviors.

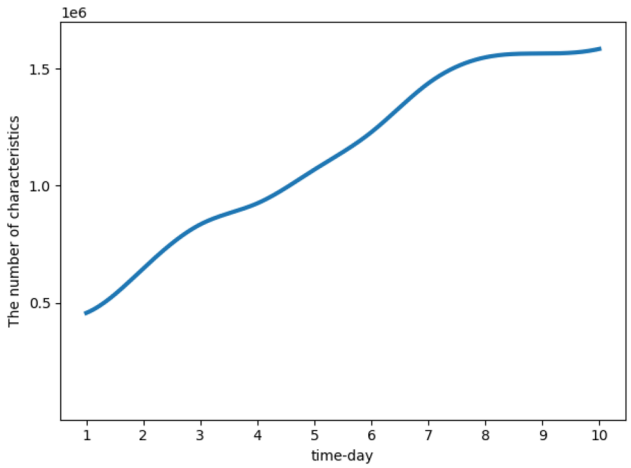

Fig. 4 Observed time—number of features

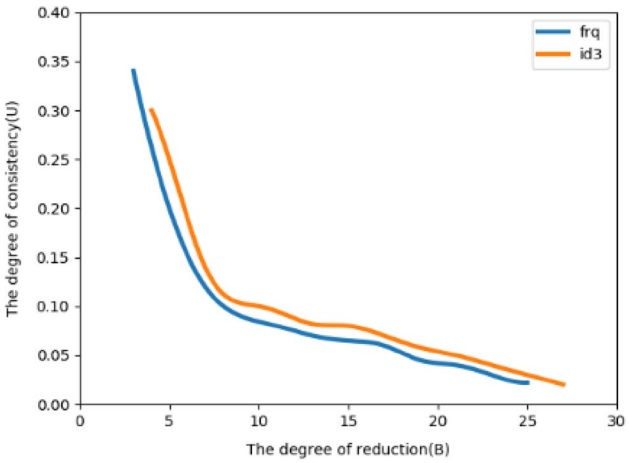

Fig. 5 Comparison of discretizing algorithms

6 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 128 GB. The algorithm is implemented in C++ and python language. The experiment first analyzes the process of feature discretization and feature stability screening, then, the comparison of our algorithm with other similar alternatives is evaluated.

6.1 Characteristics discretization analysis

The flow data of this paper is collected from the China Education and Research Network of Jiangsu Province, including three units, ten network segments and a total of "67,584" educational network IPs. "1,699,198" characteristics are obtained after 10 days continuous observation. The feature growth curve is shown in Fig. 4, where x-axis represents the observed time and y-axis is the number of characteristics.

Two parts contained in the characteristics growth curve. In the first part, the number of characteristics grows faster with the observed period is 1–7 days, while, the growth rate of the second part is relatively stable and the slowing growth trend of the number of features during this period of time (7–10 days). There are two reasons contribute to this phenomenon. On the one hand, some users' network behavior is relatively fixed with little changes in a short period of time. Hence, the observed behavior characteristics are relatively limited, and all characteristics can be observed if the size of the observed window is large enough. On the other hand, traffic behavior is periodic. From the curve, the number of features starts to slow down on the eighth day. It may be explained that the periodicity of network users' behaviors exists, so that the number of features change largely in a

single day, but traffic behaviors are similar in one cycle. In order to improve mining efficiency and reduce noise, in this paper we adopt two discretization approaches to solve the problem of overly detailed features, as shown in Fig. 5.

In the process of characteristics discretization, the parameter (R) of minimal number of samples is chosen as a termination condition. In other words, when the number of samples in a certain numerical range less than R, the numerical range will not be divided any more. ID3 algorithm based on information gain and the unsupervised FRQ algorithm are used to discretize these numerical characteristics. Three evaluation criteria of the discretization results are included, i.e., the degree of reduction (brief, B), the degree of consistency (uniformity, U) between the discretization results and original data information, and classification accuracy. Since this work does not involve classification problems, the B and the U is chosen as the evaluation criteria to verify the superiority of ID3 algorithm. B = N_value\N_symbol, where N_value is the number of numerical characteristics before discretization and N_symbol is the number of numerical characteristics after discretization. The larger B is, the stronger degree of reduction will be. U = IPsPerValue\IPsPerSymbol, where, IPsPerValue represents the number of IP address to which the sample belongs in each numerical characteristic, and IPsPerSymbol represents the number of IP address to which the sample belongs in each numerical characteristic after discretization. U is the ratio of the mean value between IPsPerValue and IPsPerSymbol. If no discretization is performed, U is equal to one, and the consistency degree is the highest. As the discretization proceeds, the U value and the consistency value is gradually decrease. With the same parameter R, the comparison results between the two algorithms are shown in Fig. 5. The consistency coefficient decreases with the increase of B as show in it, however, when setting the same value of B, ID3 is superior to FRQ as it has a high degree of consistency. In the experiment, when B is configured to 19, ID3 algorithm obtains a best discretized characteristics set.

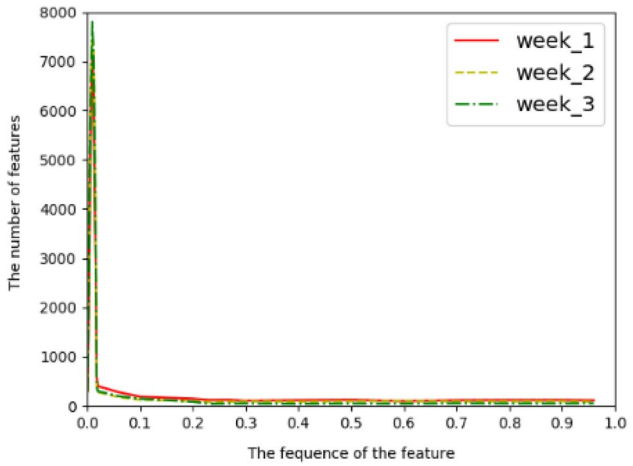

Fig. 6 Feature-frequency distribution chart for 3 weeks

6.2 Stability characteristics screening analysis

In order to determine the appropriate frequency threshold of obtained characteristics, 1 week is selected as the characteristic statistical cycle. Observing for 3 weeks continuously, the feature-frequency distribution chart is drawn, as shown in Fig. 6.

Although, there are some differences of the number of features in the features-frequency chart among the three observed cycles, they all show obvious heavy-tailed trait. Most of the traffic characteristics are located at the head of the curve, the frequency is less than 0.02 and the stability is weak, and only a small number of traffic characteristics are highly stable. In the experiment, 2% is chosen as the minimum stable occurrence frequency and the number of stable characteristics of the three cycles are 1099, 1113, and 1073, respectively.

When updating the characteristic spectrum, the rate of change with the value of 5% is used as the threshold value. If the proportion of initial characteristic account for the current spectrum more than 95%, which indicates that little changes occurred in network traffic. In this case, it is not necessary to reconstruct the IP address traffic behavior characteristic spectrum. For example, in the experiment, 1099 stable features appear in the first cycle, and the maintenance rate of the characteristic spectrum in the second cycle is 98.56%. Similarly, the maintenance rate of the characteristic

Table 11 Comparison of IP traffic behavior analysis methods

Approach	Xu et al. (2011)	Jakalan et al. (2016)	Hongbo et al. (2010)	Our work
Number of clusters	30	32	36	42
The dimension of features	Only spatial dimension with all the IPs in a managed domain	Only spatial dimension with the IPs are spilt in and out of the managed domain	Only Intensity dimension including flow size, ratio of send and received packets	Temporal, spatial, category and intensity dimension
Time consumption (day)	1.6	1.8	1.1	1.3
Memory consumption (MB)	248.4	295.6	168.2	177.8

Table 12 Some identified traffic behavior patterns of different approach

ID	Size	Flows	Outside peers	Inside peers	Destination port	Patterns
Cluster 5 in Xu et al. (2011)	75	648	0	178	8081	(sip [178], spt *, dip [75], dpt 8081)
Cluster 18 in Jakalan et al. (2016)	35	705	17	0	3389	(sip [17], spt *, dip [35], dpt 3389)

spectrum in the third cycle is 98.38%. According to the data above, we can see that the rate of change of the characteristics is less than 5%. Therefore, we only need to replace the stable features in the characteristic spectrum rather than to reconstruct it.

6.3 Comparison with other mechanisms

Table 11 compares the method proposed in our work with other three existing IP traffic behavioral methods (Xu et al. 2011; Jakalan et al. 2016; Jiang et al. 2010). Based on the characteristic spectrum technique used, the computation cost of the proposed solution is much lower than the methods of Xu et al. (2011) and Jakalan et al. (2016). In order to verify our approach is more fine-grained and comprehensive in analyzing the IP traffic behavior, we conduct an experiment with an empirical data on one network segments of 3 units including 8796 IPs in Jiangsu Education and Research Network. Firstly, different IPs are clustered based on the similarity of their traffic behavior, then, the behavior pattern of each cluster is analyzed. Seen from Table 11, our approach has more clusters than the others. The more clusters, the more detailed patterns for a group of end hosts sharing similar behaviors can be revealed. Next, we randomly select one cluster among different methods to compare the granularity and comprehensiveness of the behavior pattern, shown in Table 12.

Xu et al. (2011) and Jakalan et al. (2016) propose an approximation algorithm to discover communities on a large scale in the managed domain based on the bipartite networks and one mode projection graph. Jakalan et al. (2016) classify inside IP addresses that are connected to the same IP addresses from the outside network in one group, which is more useful than Xu et al. (2011) through providing a better understanding of what services are requested from or provided to the outside network. As both of them explored the behavior similarity in the spatial dimension, we use the same form to describe their behavioral pattern, for example, cluster 5 consists of 75 destination hosts (dip [75]) with which 178 source hosts (sip [178]) communicate via destination port 8081 (dpt [8081]) and random source ports (spt *), compared with the roles (A and B) they cloud not identify the behavior pattern of a specific application and the traffic behavior in a certain period of time, such as whether they have periodicity or rhythmicity or not. However, our proposal can analysis the traffic distribution of different applications and identify the rhythmicity of an IP activity. In a word, our approach with a fine-grained analysis than Xu et al. (2011) and Jakalan et al. (2016). Besides, Jiang et al. (2010) extracted the volume of traffic in different attributes (flow size, ratio of upload and download packets and the size distribution of flow in different applications, etc.) 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 simply be missed. Although, some metrics such as traffic distribution in time and traffic distribution in space is similar to ours, it is not as detailed as ours. In our work, some dual-attribute metrics are divided based on unit time, which could be used to describe the rate of changes of other divided attributes. There are signs demonstrate that network traffic maybe abnormal if it has large rate of change of traffic per unit time. Similarly, the traffic distribution in space dimension is also divided based on peer IP and peer IP address blocks. Not only the observed IP communicate with which end hosts, but also their communication range can be found, however, the reference in Jiang et al. (2010) doesn't have such functions. Results demonstrate that the roles mined in our work can be used to describe the traffic behavior with a fine-grained way and more comprehensiveness than previous studies.

7 Conclusion and future work

Both network operation management and network security monitoring need to discover and describe the IP address traffic behavior. An IP address role mining approach from four dimensions including temporal, spatial, category and intensity is designed, which can be applied to discover the behavior information of IP address in a large number of noise-containing flow data. Towards to this end, nine single-attribute and thirty-nine dual-attribute metrics are extracted to construct the IP address traffic characteristic spectrum, and a characteristic spectrum matched IP address role mining algorithm is also given. After the analysis of the traffic behavior profile of the network segment 202.194.*.*/20, we find that the traffic has obvious rhythmic behavior, which conforms to the habits of different users in the network.

After interpreting the roles of A and B, this work has not only discovered abnormal behaviors such as scanning, but also differentiated specific server types. Moreover, administrators can describe user behavior habits by intercepting some characteristics as needed, such as the mainstream behavioral habits of users. Therefore, the description granularity of this paper is more detailed and comprehensive compared with other literatures.

In the future, this paper intends to expand the scope of observations by conducting the discovery of IP address roles from multiple networks, such as corporate networks and wireless network. Analyzing the differences of traffic behaviors of the IPs in different network, and look for IP address roles that have universal or greater application scope. What's more, we intend to combine frequent pattern mining, IDS, and other techniques to analyze the hidden user behavior that participate in the activities within a short time but not intensively. We believe that as the expansion of the network size, the flow data will face greater storage pressure and hence the idea of role of IP address will have a broader application prospect.

Acknowledgements This work was conducted under the support of Jiangsu Key Laboratory of Computer Networking Technology and the Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, and some projects including the National Natural Science Foundation of China under Grant (No. 61602114), CERNET Innovation Project (No. NGII20170406) and Key Research and Development Program of China under Grant (No. 2017YFB0801703). 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

Chen, L., Gong, J.: Fast application-level traffic classification using NetFlow records. J. Commun. **33**(1), 145–152 (2012)

Chen, Y., Hwang, K., Ku, W.S.: Collaborative Detection of DDoS attacks over multiple network domains. IEEE Trans. Parallel Distrib. Syst. **18**(12), 1649–1662 (2007)

Cui, S., Li, W., Yi, L., Li, C., Zhu, L., Jiang, Z.: A bibliometrical analysis of status on animal behavior in China. Acta Theriol Sin **36**(4), 476–484 (2017)

Fayaz, S.K., Tobioka, Y., Sekar, V.: Bohatei: flexible and elastic DDoS defense. In: USENIX, pp. 817–832 (2015)

Gañán, C., Cetin, O., van Eeten, M.: An empirical analysis of ZeuS C&C lifetime. In: Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security (ASIA CCS '15), pp 97–108. ACM, New York (2015)

Gong, J., Zang, X.D., Su, Q., Hu, X.Y., Xu, J.: Survey of network security situation awareness. J. Softw. **28**(4), 1010–1026 (2017)

Grill, M., Nikolaev, I., Valeros, V., Rehak, M.: Detecting DGA malware using NetFlow. In: IFIP/IEEE International Symposium on Integrated Network Management, pp. 1304–1309 (2015)

Himura, Y., Fukuda, K., Cho, K., Borgnat, P., Abry, P., Esaki, H.: Synoptic graphlet: bridging the gap between supervised and unsupervised profiling of host-level network traffic. IEEE/ACM Trans. Netw. **21**(4), 1284–1297 (2013)

Hoque, N., Bhattacharyya, D.K., Kalita, J.K.: Botnet in DDoS attacks: trends and challenges. IEEE Commun. Surv. Tutor. **17**(4), 2242–2270 (2015)

Iliofotou, M., Gallagher, B., Eliassi-Rad, T., Xie, G., Faloutsos, M.: Profiling-by-association: a resilient traffic profiling solution for the internet backbone. In: Proceedings of the 6th International Conference, Philadelphia, Pennsylvania (2010)

Jakalan, A., Jian, G., Zhang, W., Qi, S.: Clustering and profiling ip hosts based on traffic behavior. Comput. Netw. **100**, 99–107 (2016a)

Jakalan, A., Gong, J., Su, Q., Hu, X., Abdelgder, A.M.S.: Social relationship discovery of IP addresses in the managed IP networks by observing traffic at network boundary. Comput. Netw. **100**, 12–27 (2016b)

Jiang, H., Ge, Z., Jin, S., Wang, J.: Network prefix-level traffic profiling: characterizing, modeling and evaluation. Comput. Netw. **54**(18), 3327–3340 (2010)

Karagiannis, T., Papagiannaki, K., Faloutsos, M.: BLINC: multilevel traffic classification in the dark. Proc. ACM SIGCOMM **35**(4), 229–240 (2005)

Kheir, N.: Behavioral classification and detection of malware through HTTP user agent anomalies. J. Inf. Secur. Appl. **18**(1), 2–13 (2013)

Kozik, R.: Distributing extreme learning machines with Apache Spark for NetFlow-based malware activity detection. Pattern Recognit. Lett. **101**, 14–20 (2018)

Krishna Reddy, P., Kitsuregawa, M., Sreekanth, P., Srinivasa Rao, S.: A graph based approach to extract a neighborhood customer community for collaborative filtering. Databases Netw. Inf. Syst. **2544**, 188–200 (2002)

Deri, L.: Open source VoIP traffic monitoring. http://131.114.21.22/VoIP.pdfS.Retrieved Accessed 3 June 2012

Lee, D.J., Brownlee, N., Host measurement of network traffic. In: Telecommunication Networks and Applications Conference, pp. 282–287 (2007)

Li, B., Springer, J., Bebis, G., Gunes, M.H.: A survey of network flow applications. J. Netw. Comput. Appl. **36**(2), 567–581 (2013)

Marnerides, A.K., Schaeffer-Filho, A., Mauthe, A.: Traffic anomaly diagnosis in Internet backbone networks: a survey. Comput. Netw. **73**, 224–243 (2014)

Miao, L.H., Ding, W., Yang, W.: Extracting and analyzing internet background radiation in live networks. J. Softw. **26**(3), 663–679 (2015)

Saied, A., Overill, R.E., Radzik, T.: Detection of known and unknown DDoS attacks using Artificial Neural Networks. Neurocomputing **172**, 385–393 (2016)

Schatzmann, D., Mühlbauer, W., Spyropoulos, T., et al.: Digging into HTTPS: flow-based classification of webmail traffic. In: 10th ACM SIGCOMM Conference on Internet Measurement, pp 322–327 (2010)

Umer, M.F.S., Yaxin, M.B.: Flow-based intrusion detection: techniques and challenges. Comput. Secur. **70**, 238–254 (2017)

Umer, M.F., Sher, M., Bi, Y.: Flow-based intrusion detection: techniques and challenges. Comput. Secur. **70**, 238–254 (2017)

Wei, S., Mirkovic, J., Kissel, E.: Profiling and clustering internet hosts. In: Proceedings of the International Conference on Data Mining, pp. 269–275 (2006)

Weijie, G.: The parallel and implementation for network behavior observations system. (MS. Thesis), Southeast University, pp. 3–20 (2010)

Xiao, J.Q., Wang, D.: Construction of behavioral spectrum of the Yangtze finless porpoise in captivity. Acta Hydrobiol. Sin. **29**(03), 253–258 (2005)

Xu, K., Wang, F., Gu, L.: Network-aware behavior clustering of internet end hosts. In: INFOCOM, pp. 2078–2086 (2011)
Xu, K., Wang, F., Gu, L.: Behavior analysis of internet traffic via bipartite graphs and one-mode projections. IEEE/ACM Trans. Netw. **22**(3), 931–942 (2014)
Zhao, D., Traore, I., Sayed, B., Lub, W., Saad, S., Ghorbani, A., Garant, D.: Botnet detection based on traffic behavior analysis and flow intervals. Comput. Secur. **39**, 2–16 (2013a)
Zhao, D., Traore, I., Sayed, B., Wei, L., Saad, H., Ghorbani, A., Garant, D.: Botnet detection based on traffic behavior analysis and flow intervals. Comput. Secur. **39**, 2–16 (2013b)
Zheng, D.L.: Behavioral ecologic research on several kinds of Africa herbivores in semi-nature. [Master Theisi], Shandong Normal University, pp. 3–15 (2005)

Xiaodong Zang is a Ph.D. Candidate in 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. In 2013, he received his MSc in Computer Science and Technology from Nanjing University of Posts and Telecommunications, China.

Jian Gong is a professor in the School of Cyber Science and Engineering, Southeast University. His research interests are network architecture, network intrusion detection, and network management. He has received his BS in computer software from Nanjing University, and his PhD in computer science and technology from Southeast University.

Siyi Huang received hers M.S. from School of Computer Science and Engineering, Southeast University, Nanjing, China. Hers research interests include computer networks and security.

Xiaoyan Hu is currently an assistant professor in School of Computer Science and Engineering, Southeast University, Nanjing, China. She received her Ph.D. from Southeast University in 2015. Her research interests include future network architecture and network security.

Yun Yang is currently a master degree candidate in School of Computer Science and Engineering, Southeast University, Nanjing, China. His research interests include computer networks and security.