Profiling IP Hosts Based on Traffic Behavior

Ahmad Jakalan, Jian Gong, Shangdong Liu
School of Computer Science & Engineering, Southeast University
Jiangsu Key Laboratory of Computer Network Technology
Nanjing, China
{ahmad, jgong, sdliu }@njnet.edu.cn

***Abstract*— Internet is growing very fast in size and applications which causes more complexity in its structure which demands more efforts for monitoring and management. For an efficient network security and management it's required to have a better understanding of its structure and traffic caused by different elements (IP hosts). There is a need to improve systems and methods that can provide this kind of knowledge and understanding. The objective of this research is to study the behavior of IP Network nodes (IP hosts) from the prospective of their communication behavior patterns to setup hosts' behavior profiles of the observed IP nodes by clustering hosts into groups of similar communication behaviors. There are many potential applications of this work; the results of this research will be useful to the network management and Network Security Situational Awareness (NSSA) in addition to its applications in studying the network user's behavior.**

Keywords- Computer Networks Security, Host Profiling, IP Networks, Traffic Behavior.

I. INTRODUCTION

IP networks Host behavior profiling refers to observing measured flow data from Internet backbone and extracting information which is representative of the communication behavior or usage patterns of the observed hosts. It is useful in understanding the behavior of the monitored network and in deriving guidelines of normal and abnormal activities within that context. IP Profiling at a large scale faces several challenges like the huge number of active hosts observable in the backbone traffic flows and the sporadically appearance of the observed hosts. Host profiling and clustering aims at identifying dominant and persistent hosts behaviors and creating groups with similar behaviors, this is very useful for many applications of Internet security such as Network Security Situational Awareness NSSA, DDoS defense, worm and virus detection, botnet detection, etc. For example worm infection or any attack on the network might cause a sharp change in the host's behavior, so detecting attacks on the network will be easier if we can profile hosts behaviors so that sharp changes in hosts' behaviors will be detected. Our study is based on China Education and Research Network (CERNET) backbone data. We use IP Flow data collected from Netflow of border routers generated by over different periods of time. The collected data is stored in files of a limited period of 5-minutes to be used later for analysis. This study is based on CERNET backbone data, but the method could be applied on general Internet traffic analysis. This paper is organized as follows: Section 2 reviews a number of related works. In section 3 we explain our methodology briefly. The extraction of the most significant IP addresses to be profiled is described in section 4, and then in section 5 we present the selection and extraction of communication pattern features, and the results with the discussion are presented in section 6, and then the final conclusion is in section 7.

II. RELATED WORKS

Different researches appeared for profiling Internet traffic for different purposes, detecting network traffic anomalies was the main purpose of most of them. BLINC[1] identify application footprints in traffic streams by classifying traffic flows according to the applications that generated them. CAI Jun et al. [2] measures the dynamic changes of host communities for the purpose of anomalous detection. Xu Kuai et al. [3, 4] identify common traffic profiles as well as anomalous behavior patterns based on behavior profiles of Internet backbone traffic in terms of communication patterns of end-hosts and services. Xu Kuai et al. [5] characterize the behavior of the significant clusters and groups the clusters into classes with distinct behavior patterns. Vanessa F et al. [6] identify anomalous behavior where the behavior of a host raises an alert only when a group of host profiles with similar behavior (cluster of behavior profiles) detect the anomaly, rather than just relying on the host's own behavior profile to raise the alert. Different techniques used in profiling IP nodes; Xu Kuai et al. [7, 8] applied spectral clustering algorithms on the one-mode projection of bipartite graphs to find the clustered behaviors of end hosts in the same network prefixes. Unsupervised data mining techniques were applied also for profiling end nodes[9, 10], Guillaume D et al. [9] applied minimum spanning tree (MST) clustering technique. Karagiannis et al. [10] build and continuously update activity graphlets that capture all the current flow activity, and then compress to retain a profile graphlet. The infinite dimension of graphlets make it difficult to apply unsupervised clustering in addition to that only simple patterns can be identified while neither new class nor any mixture of traffic can be discovered. Songjie Wei et al. [11] applied Dice similarity function to calculate the similarity of hosts' communications to create profiles and then used hierarchical clustering techniques on the profiles to build a

dendrogram containing all the hosts. The user behavior networks that connect users with servers across the Internet were studied in [12] to classify the clients into normal and abnormal communities. Many other works exist on profiling Internet backbone traffic [13-17] for profiling and classifying endpoints characteristics by extracting the information about endpoints from elsewhere using collected and combined information freely available on the Web. It is well-known that the Internet traffic is heavy-tailed, most significant clusters will dominant the traffic behavior, so that the paper will concentrate on the behavior profiling of these most significant IP addresses. Previous researches focus on profiling Internet hosts over short periods of time (1~5 minutes) which is considered a relatively short periods to setup host behavior profiles, but in our study we analyze behavior patterns over a long period of time (one hour) to be able to setup a more stable host behavior profiles.

III. METHODOLOGY

The main purpose to study the behavior of a single IP address is to be able to setup a profile of the IP addresses. The problem here is how to define the details of these profiles and which metrics needed. The content of this profile should be selected carefully to help the further work. The most important points should be considered when building this profile includes the data structure and the content of the profile, and how often it should be updated.

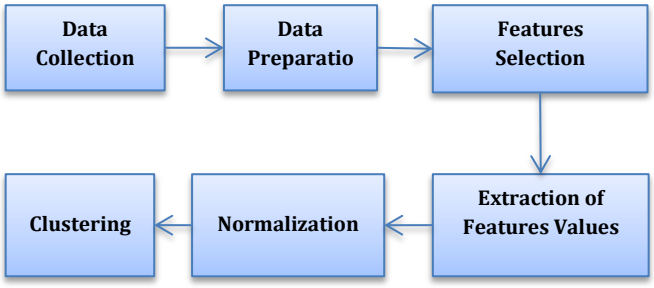


Figure 1 A schematic process of hosts behavior clustering

Because it is not reasonable to setup a profile for each observed IP address, so they are clustered. Clustering of IP profiles will be based on their network traffic behavior patterns to identify the service behind this IP host. Individual host's behaviors could change over time but the profile of a legitimate host tends to fall into the same category for a moderately long time. Grouping hosts into categories is useful to build models of legitimate Internet communications. These models will be useful in the detection of suspicious changes in the backbone traffic, which are usually a sign of an Internet-wide security problem. An accurate categorization of Internet hosts can help identify malicious Internet hosts (and their users) from the mass of legitimate ones. Machine learning will be applied for clustering profiles. For machine learning approaches, feature selection is very important and needs to be specific to the problem. A combination of features will be used, some of them are directly extracted, and others are calculated using simple calculations or statistical analysis or obtained after applying techniques from the information theory. It's not possible to study all IP addresses or all clusters obtained, so the attention of study will be focused on the most significant IP nodes that initiate most of the observed traffic, and we want to study them over a long enough period of one hour to get more representative and reliable profiles.

IV. EXTRACTION OF THE MOST SIGNIFICANT IP ADDRESSES

It's not possible to profile every IP address appears over the internet, even each IP address in the trace, so we focus on the most significant IP addresses. We adopted a cost effective method to extract the most active IP addresses that initiate most of the flows in the trace. We have found that excluding 10% of the total flows will reduce the number of IP addresses that need to be analyzed in a very efficient way. Fig.1 shows the number of significant SrcIPs, DstIPs from the total and distinct number of IP addresses observed in the trace of one day within periods of one hour, and because our study focuses on active flows initiated by the IP address, so we extracted the most significant SrcIPs.

Let n denotes to total number of flows, m is the number of distinct elements of srcIPs, If $X = \{x_1, x_2, \dots, x_m\}$ is the complete list of distinct SrcIPs, let $P(x_i)$ denotes the possibility of appearance of x_i in the flows of the trace during the period of study. We select an epsilon value $\varepsilon = 0.1$ to exclude the srcIPs that initiate flows less than 10% of the total flows, and analyze IP addresses that initiate more than 90%. The remaining significant srcIPs S is the list of SrcIPs that initiate flows more than 90% of the total flows

$$S = \{x \mid \sum p(x_i) > 1 - \varepsilon\} \quad (1)$$

Figure 2 shows a 10 base log scaled curves to demonstrate the relation between the total number of flows(blue curve) with the total number of distinct Src/Dst IP addresses over a duration of complete one day with 24 periods of one hour. The number of distinct Src/Dst IP addresses that appear in 90%, 80% of the total traffic captured we called them the most significant IP addresses

We may notice that for these periods over one day, the maximum number of flows per hour may reaches tens of millions with about one million of different source IP addresses. If we exclude 10% of the total traffic captured we may get a list ten times less than the original list of SrcIPs that initiate 90% of the total traffic captured by Netflow. For our study, to get a more reliable and more reasonable results we have excluded 10% of flows and studied the 10% of SrcIPs that initiate more than 90% of the total traffic.

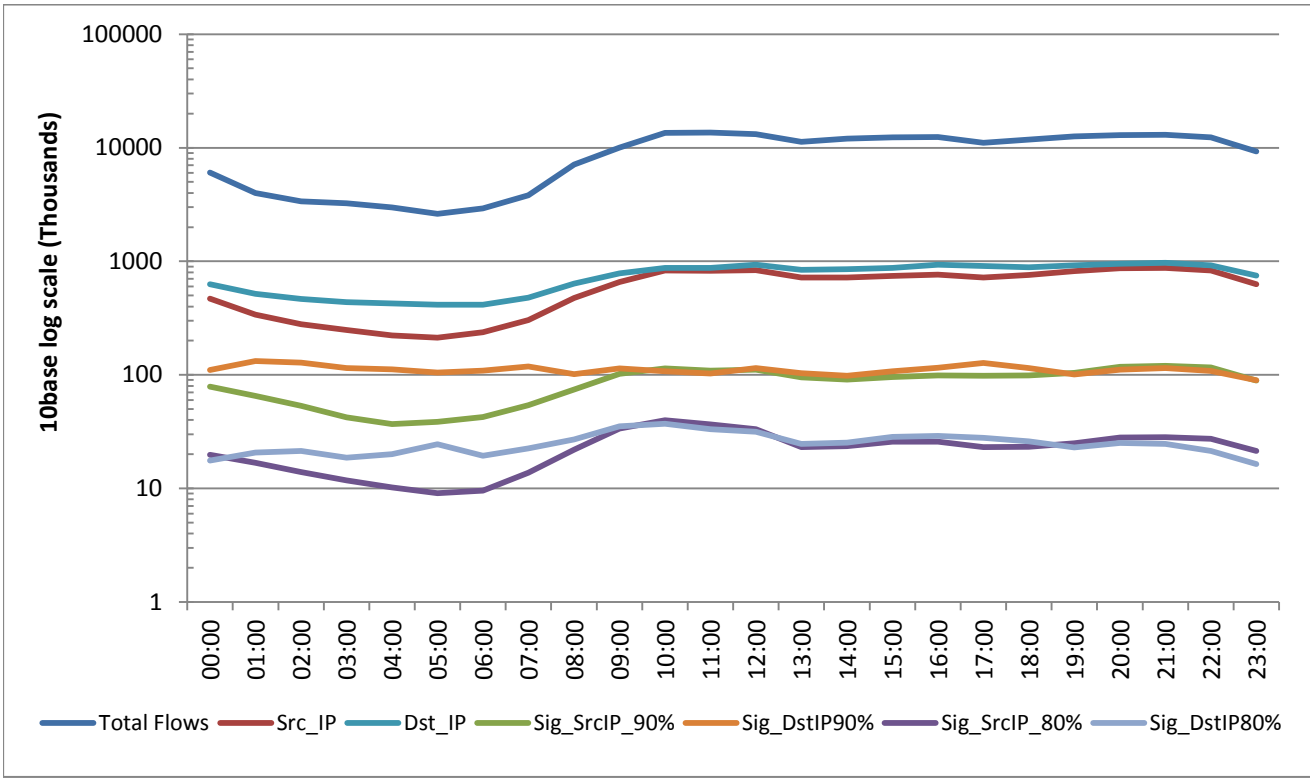


Figure 2 a graph showing 10 base log scaled curves to demonstrate the relation between the total number of flows(blue curve) with the total number of distinct Src/Dst IP addresses over a duration of complete one day with 24 periods of one hour. The number of distinct Src/Dst IP addresses that appear in 90%, 80% of the total traffic captured we called them the most significant IP addresses

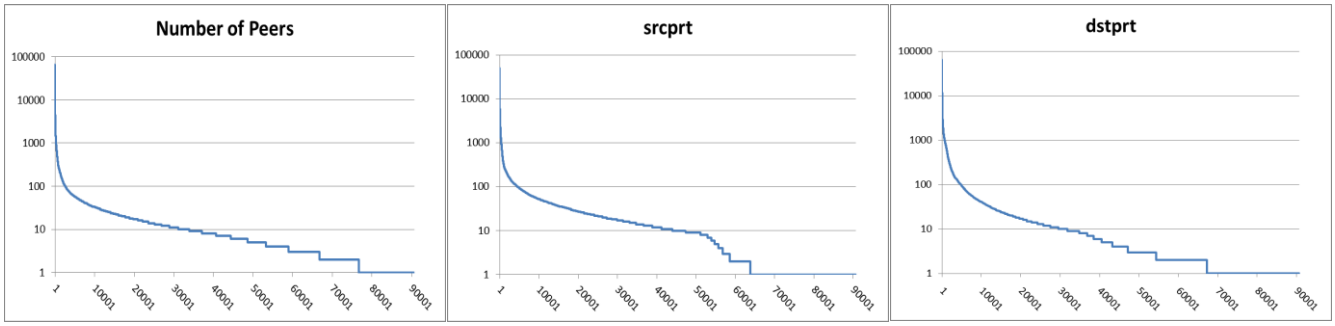
V. SELECTION AND EXTRACTION OF COMMUNICATION PATTERN FEATURES

For the efficiency of processing and ease of interpretation we need to keep the number of feature space as low as possible, but on the other side to allow the discrimination of different host behaviors it should present host behavior carrying rich enough information. Our focus will be on active communication initiated by the profiled host. We found the following features are the most important to represent host behavior communication patterns:

1. Number of peers (or the count of unique Destination IP addresses): IP addresses to which at least one packet is sent from this IP. This feature distinguishes the host community of peers that receive traffic from this IP. The importance of this feature comes from that this feature distinguishes one-to-one communications (like P2P or downloads) from one-to-several (web browsing) and one-to-many (netscans). Actualy it represents the social popularity of this IP address. Figure 3 (a) demonstrate the distribution in the number of peers over a duration of one hour, we may notice that a very little number of IP addresses have a very big number of peers while most of them send traffic to less than 10 peers.
2. The ratio of the entropy of the first byte of DstIP to the entropy of the fourth byte of DstIP H(IP1)/H(IP4).
3. The ratio of the entropy of the second byte of DstIP to the entropy of the fourth byte of DstIP H(IP1)/H(IP4).
4. The ratio of the entropy of the third byte of DstIP to the entropy of the fourth byte of DstIP H(IP1)/H(IP4).

These features reflect the social role of a host, they characterize the dispersion observed in the list of peers (or Destination IPs) associated with a SrcIP. Distribution of peers over the IP space is not random in real cases, the first and second bytes usually correspond to locations or ISPs, while the third one correspond to companies or organizations, while the fourth one represents hosts in the same sub-network. Most regular traffic entropy measured on the second and the third bytes tend to be just a little lower than that on the third and the fourth, so a large difference in these entropies is likely to betray scanning.

5. The ratio of the number of source ports per the number of peers: servers usually receive requests on a single port, and use the predefined specific port as a source port in the response for classical protocols, while clients usually open a different random port for each connection to a server. Figure 3 (b) shows the number of distinct source ports, we may notice that a small number of hosts use a very big number of



(a) (b) (c)

Figure 3 (a) shows number of peers of each SrcIP, (b) shows number of distinct source ports used by each IP address to communicate with otthers, (c) shows number of distinct destination ports

source ports to send traffic to others, about half of the senders use from 10 to 100 distinct source ports, and we may notice that about 25% of the hosts use a single port which could be a server behavior if they have heavy traffic or just a single appearance of a normal user activity.

6. The desperation of distribution in source ports: The number of distinct source ports itself may not reflect valuable meaning, for example a host providing a web service on port 80 will use this port to send http traffic, at the same time it may be using other ports for traffic of other different services, but for example mostly it is using port 80, when using only the distinct number of ports, this port will be represented as a one value and will not reflect the frequency of using this port while entropy of ports represent the frequency of used ports.
7. The ratio of the number of destination ports per the number of peers reflects the role of the host. Scanning open ports on a single or some IP addresses will result a high value of this feature, while a very low value may represent a scan of a single port on many IP addresses. Figure 3 (c) shows the number of distinct destination ports, we may notice that a small number of hosts use a very big number of distinct destination ports on their peers to send traffic to them, about half of the receivers use from 10 to 100 distinct destination ports, and we may notice that about 25% of the hosts use a single destination port which could be a server behavior receiving requests (DNS or HTTP) requests.
8. The desperation of distribution in destination ports: similar to feature number 6 also this feature reflects the distribution of destination ports.
9. The mean number of packets per flow distinguishes elephant flows from mice flows (non-connected flows and could be an attack).
10. The mean packet size: small-size packets mostly consist of signaling or scan traffic while large-size packets are data exchange.
11. The mean number of flows per peer reflects consistency of traffic between these two hosts. While the flow is created by Netflow over a specific period of time, so new flow is created to the same destination IP address if the connection stays active for a period longer than Netflow's predefined period of the flows.
12. Mean duration of flow differentiate between connected vs. non-connected flows which is possible to be attacks.
13. The entropy of protocols used by this IP to communicate with other IP addresses distinguishes service providers that mostly use single protocol from other clients that may use different protocols. IP protocol value comes from the flow record (where 6=TCP, 17=UDP).
14. The entropy of type of application distinguishes service providers from clients that normally use more applications simultaneously. The type of application comes from the flow record (where FTP=1, www=2, Mail=3, P2P=4, Service=5, Interactive=6, Multimedia=7, Voice=8, others=0).
15. Number of sent SYN-ACK: When two computers attempt to communicate they negotiate the parameters of the network TCP socket connection before transmitting data, in all situations the service provider whose service is requested should send the SYN-ACK message when it accepts the request of clients to start or end the session.

VI. Clustering Results and Discussion

The selected features values vary with large ranges like number of peers, and vary in narrow ranges like entropies. So the values of features need to be normalized to get values within the range [0, 1] by dividing each feature on the maximum value of the feature. As we have mentioned that we do not have advance knowledge of the exact number of host categories we are going to create and of the defining features of each category, and the number of elements in

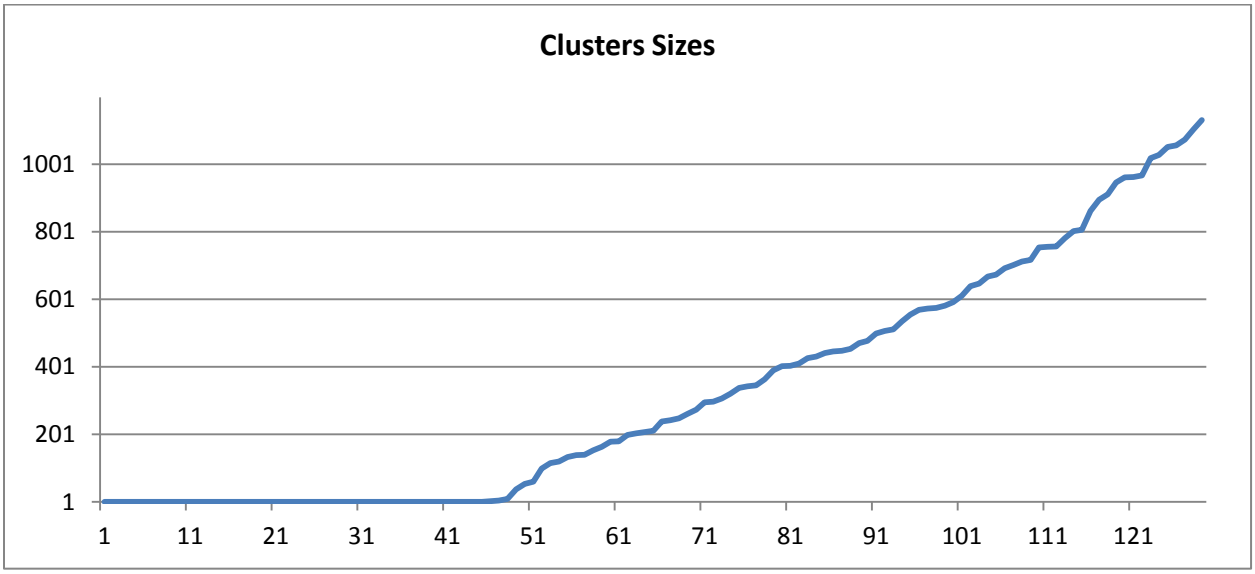


Figure 4 shows a curve of clusters sizes, the first 46 clusters with size of 1 element are the outliers which they don't belong to any clusters. We may notice that a small number of clusters with sizes between 2 and 100 elements while most of the clustrs are bigger than 100

each category, the clustering techniques in data mining come as an appropriate tool for host classification. In our research we have applied DBSCAN[18] clustering algorithms implemented by weka[19] on the extracted features space to cluster hosts based on behaviors, clusters were labeled with numbers. Figure 4 shows a curve of clusters sizes, the first 46 clusters with size of 1 element are the outliers. We may notice that a small number of clusters with sizes between 2 and 100 elements while most of the clusters are bigger than 100. We have selected some clusters and went back to their traffic behaviors, we found that it's possible to notice some significant clusters like those presented in Table 1:

A. Scanning a single port

The cluster labeled with number 6 the number of elements in this cluster is not big 17 SrcIPs. We may notice the big number of peers, and the small size of packets, no SYN-ACK signals sent from these IPs, a single source port were used in transmission to a single destination port on the destination IPs. A single packet is sent in each flow from the SrcIP with a very low duration of flow. All SrcIPs in all of

TABLE 1: some selected behavior clusters (values here are the real values not the normalized)

Cluster label	**1**	**3**	**6**	**9**	**13**
Number of elements	234	803	17	3	132
Averages of the values of extracted Features of the cluster					
Number of peers	1	3	1549	1	2343
H_IP1/4	0	0.072	0.271	0	0.364
H_IP2/4	0	0.070	0.401	0	0.447
H_IP3/4	0	0.083	0.995	0	0.805
Number of srcprts per peers	50	0.891	0.001	1	0.0242
H_srcprt	4	0.067	0	0	0.005
Number of dstprts per peers	1	31	0.00084	9431.05	6.38
H_dstprt	0	4.117	0	9.704243	8.34
Mean pkts per flow	1.2	440	1	1.15079	2
Mean pkt size (byte)	590	1454	75	483	1225
Mean flows per peer	56	44	1	12005.47	8
Mean duration of flow (ms)	6526	14695	0.0006	0.000348	4559
H_prot	0	0	0	0	0.0004
H_toa	0.0027	0.008	0	0	0.0053
Number of SYN-ACKs	0.0256	2.42	0	26	636

their transmission used only one protocol and a single type of application, and a single flow is made to DstIPs. We may notice also that the changes in the third and fourth bytes of DstIPs is much bigger than the changes in the first two bytes, we may notice that the change in the third destination IP is very slightly lower that of the fourth byte which means a scan over class B.

B. Port Scanning on a single host

The cluster labeled with number 9 the number of elements in this cluster is not big 3 SrcIPs. We may notice that they communicate with a single destination host with a big number of ports per peers, and a very high value of entropy on the destination port value, a small size of packets, a single source port were used in transmission to a very large number of destination ports on the destination IPs. A single packet is sent in each flow from the SrcIP with a very low duration of flow. All SrcIPs in all of their transmission used only one protocol and a single type of application, and it seems that a single flow is made to new different destination port on the DstIP, and since they communicate with single destination IPs so that the value of entropy on the destination IPs is none noticeable.

C. Server traffic behaviour

The cluster with label 13 includes 132 elements, they show a server traffic behavior, they send traffic to a very high number of peers (clients in this situation) with a very low entropy of source ports and the maximum entropy of destination ports which means the change in the ports on the servers is very low while the changes in the ports on clients is very high (a new dstPrt port for each connection). We notice that the hosts in this cluster send a big number of SYN-ACK signals which can't be sent from the host that initiate a connection (client) but can be sent from the hosts that provide a service to other clients here we call them as servers. Also we may notice that the packets transmitted are medium in size not small and not big which means a normal traffic and a medium duration of flows. The changes in protocol and type of application are very low.

D. Clients sending http like requests

The size of cluster with the label 1 is medium with 234 hosts each host is transmitting to a single destination flows with a small packet-size but a slightly long duration of flows more than the duration required to send in average two packets with a medium to small packet size. We may notice that each host in this cluster is sending the packets to a single port on the receiver, using a different source port per flow, and also they tend to use a single protocol and a single type of application, so we may say that the hosts within this cluster are clients each one is requesting a service from a single server under a single protocol and a single application which may be http request.

E. P2P Traffic

Cluster with label 3 is considered to be relatively a big cluster of hosts initiating big traffic with a small number of peers, we may notice that the packet size tend to be so big and the number of packets per flow is also very big with a very long duration of flows and a big number of flows per destination IP, a single type of protocol and a single type of application with a relatively small number of SYN-ACK equals to the number of peers. This form of traffic is similar to that of P2P traffic.

VII. CONCLUSION

The contribution of this paper includes: (1) the selection of most important features or host behavior communication patterns to be utilized in clustering to characterize accurately and efficiently groups of host behavior traffic. (2) We presented an algorithm to extract most significant IP nodes to be analyzed instead of analyzing the complete list of millions of IP nodes that exist in the trace, this algorithm is based on the frequency of appearance of IP addresses in the flow records to study IP addresses that initiate more than 90% of the overall traffic captured. 3) We analyzed IP nodes traffic behavior on a relatively long period of traces, which help to extract a more stable host's behavior profiles. While previous studies focus only on host behavior for a relatively short period of 5 to 15 minutes, we extracted host's behavior patterns over a period of one hour which needs big data analysis to provide results in a reasonable time. Unsupervised machine learning techniques were used to cluster hosts based on their traffic patterns. Finally we selected some of the clusters and based on intuitive experience we labeled clusters.

ACKNOWLEDGEMENT(S)

This work was supported by Jiangsu Key Laboratory of Computer Networking Technology, Southeast University. The research was sponsored by National Grand Fundamental Research 973 program of China Grant No. 2009CB320505, the National Nature Science Foundation of China Grant No. 60973123.

REFERENCES

[1] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, "BLINC: multilevel traffic classification in the dark." pp. 229-240.

[2] W. X. Liu, and J. Cai, "A New Method of Detecting Network Traffic Anomalies," *Applied Mechanics and Materials*, vol. 347, pp. 912-916, 2013.

[3] K. Xu, Z.-L. Zhang, and S. Bhattacharyya, "Profiling internet backbone traffic," *ACM SIGCOMM Computer Com. Review*, vol. 35, no. 4, pp. 169, 2005.

[4] K. Xu, Z.-L. Zhang, and S. Bhattacharyya, "Profiling internet backbone traffic: Behavior models and applications," *Computer Communication Review*. pp. 169-180.

[5] X. Kuai, Z. Zhi-Li, and S. Bhattacharyya, "Internet Traffic Behavior Profiling for Network Security Monitoring," *IEEE/ACM Transactions on Networking*, vol. 16, no. 6, pp. 1241-1252, 2008.

[6] V. Frias-Martinez, S. J. Stolfo, and A. D. Keromytis, "Behavior-profile clustering for false alert reduction in anomaly detection sensors," *Proceedings - Annual*

Computer Security Applications Conference, ACSAC. pp. 367-376.

[7] K. Xu, F. Wang, and L. Gu, "Network-aware behavior clustering of Internet end hosts." pp. 2078-2086.

[8] K. Xu, F. Wang, and L. Gu, "Behavior Analysis of Internet Traffic via Bipartite Graphs and One-Mode Projections," *IEEE/ACM Transactions on Networking*, pp. 1-1, 2013.

[9] G. Dewaele, Y. Himura, P. Borgnat, K. Fukuda, P. Abry, O. Michel, R. Fontugne, K. Cho, and H. Esaki, "Unsupervised host behavior classification from connection patterns," *International Journal of Network Management*, vol. 20, no. 5, pp. 317-337, 2010.

[10] T. Karagiannis, K. Papagiannaki, N. Taft, and M. Faloutsos, "Profiling the end host," *Passive and Active Network Measurement*, pp. 186-196: Springer, 2007.

[11] S. Wei, J. Mirkovic, and E. Kissel, "Profiling and Clustering Internet Hosts," *DMIN*, vol. 6, pp. 269-75, 2006.

[12] J. L. Liu, and J. Cai, "Complex Network Community Structure of User Behaviors and Its Statistical Characteristics," in Proceedings of the 2011 Third Intl. Conference on Multimedia Information Networking and Security, 2011, pp. 366-370.

[13] Y. Jin, N. Duffield, J. Erman, P. Haffner, S. Sen, and Z.-L. Zhang, "A Modular Machine Learning System for Flow-Level Traffic Classification in Large Networks," *ACM Trans. Knowl. Discov. Data*, vol. 6, no. 1, pp. 1-34, 2012.

[14] M. Iliofotou, B. Gallagher, T. Eliassi-Rad, G. Xie, and M. Faloutsos, "Profiling-By-Association: a resilient traffic profiling solution for the internet backbone," in Proceedings of the 6th International COnference, Philadelphia, Pennsylvania, 2010, pp. 1-12.

[15] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci, "Googling the internet: profiling internet endpoints via the world wide web," *IEEE/ACM Trans. Netw.*, vol. 18, no. 2, pp. 666-679, 2010.

[16] I. Trestian, S. Ranjan, A. Kuzmanovi, and A. Nucci, "Unconstrained endpoint profiling (googling the internet)." pp. 279-290.

[17] R. Erbacher, S. Hutchinson, and J. Edwards, "Web traffic profiling and characterization," *ACM International Conference Proceeding Series*.

[18] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise." pp. 226-231.

[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10-18, 2009.