Clustering and Profiling IP Hosts Based on Traffic Behavior

Ahmad Jakalan, Jian Gong, Zhang Weiwei, Qi Su
School of Computer Science & Engineering, Southeast University, Nanjing 210096, China
Jiangsu Key Laboratory of Computer Network Technology, Nanjing 210096, China
{ahmad, jgong, wwzhang, qsu}@njnet.edu.cn

Abstract—**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 clusters of similar communication behaviors. The problem of IP address behavior analysis and profile establishment is the one that not fully discussed and the results achieved are not good enough, there is no complete solution yet. There are many potential applications of this work, the results of this research will be useful to the network management and Network security situation awareness in addition to the applications in studying the network user behavior. The contribution of this paper includes: 1) discussion about the features or host behavior communication patterns to be utilized in hosts 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. 3) We analyzed IP nodes traffic behavior on relatively long periods of traces, which help to extract a more stable host's behavior. While previous studies focus only on host behavior for relatively short periods of 5 to 15 minutes, we extract host's behavior patterns over a period of one hour which needs big data analysis to provide results in a reasonable time.**

Index Terms—Computer Networks, Host behavior profiling, Network security, traffic profiling.

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. Profiling can be done at four levels: user level, application level, host level, and network level. 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. This study is based on CERNET backbone data, but the method could be applied on general Internet traffic analysis.

The remaining of this paper is organized as follows: Section 2 reviews a number of related works. The data sources used in this study are explained in section 3. Section 4 presents some essential background and in section 5 we presented our methodology. The Selection and extraction of communication pattern features is explained in section 6 while the results and discussion are presented in section 7 then the final conclusion.

II. RELATED WORKS

Different researches has appeared for profiling Internet traffic for different purposes, detecting network traffic anomalies was the main purpose of most of them. Xu Kuai et al. [1, 2] presented a methodology for building comprehensive behavior profiles of Internet backbone traffic in terms of communication patterns of end-hosts and services to identify common traffic profiles as well as anomalous behavior patterns based on four-dimensional feature space consisting of srcIP, dstIP, srcPrt and dstPrt. CAI Jun et al. [3] measures the dynamic changes of host communities for the purpose of anomalous detection. Xu Kuai et al. [4] characterize the behavior of the significant clusters and groups the clusters into classes with distinct behavior patterns to automatically discover significant behaviors of interest from massive traffic data to help network operators in understanding and quickly identifying anomalous events with a significant amount of traffic. Vanessa F et al. [5] 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. Application identification was also one of the main purposes of these researches such as in BLINC[6] which identifies application footprints in traffic streams by classifying traffic flows according to the applications that generated them. Understanding the structure and dynamics of the user behavior networks also was an objective of some researches such as the work of Jing L et al. [7] where they analyze the structure characters and the community of the user behavior networks that connect users with servers across the

doi:10.4304/jnw.10.2.99-107

Internet, they classified the clients into normal and abnormal communities. Different techniques has been used in profiling IP nodes; Xu Kuai et al. [4] introduced an entropy-based approach to characterize the behavior of the significant clusters. In [8, 9] Xu Kuai et al. Used bipartite graphs and one-mode projection graphs to model host communication patterns observed on Internet backbone links and then 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. Data mining techniques, particularly clustering and visualization were applied widely to aid analysis of the data to identify changes in behavior of hosts. Unsupervised data mining techniques were applied also for profiling end nodes[10, 11], Guillaume D et al. [10] applied minimum spanning tree (MST) clustering technique on nine dimensional feature space evaluating host Internet connectivity, dispersion and exchanged traffic content. Graphlets has been used by Karagiannis et al. [11] to build and continuously update activity graphlets that capture all the current flow activity, and then compress them to retain a profile graphlet. The drawback in using graphlets is in the infinite dimension of graphlets which make it difficult to apply unsupervised clustering in addition to that only simple patterns can be identified while neither new classes nor any mixture of traffic can be discovered. hierarchical clustering techniques were used by Songjie Wei et al. [12] who has applied a dice similarity function to calculate the similarity of hosts' communications to create profiles of frequently-seen hosts and then used hierarchical clustering techniques on the profiles to build a dendrogram containing all the hosts, but still a level of cutting clusters into separated clusters is required which dendrograms doesn't support. Researchers has used a verity of data sources to build hosts profiles has been 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 host clusters.

III. DATA SOURCES

We use IP Flow data collected from Netflow which is an embedded instrumentation within Cisco IOS Software to characterize network operation[18]. An IP Flow is based on a set of IP packet attributes like IP source and destination addresses, Source and Destination ports... All packets with the same source/destination IP address, source/destination ports, protocol interface and class of service are grouped into a flow and then packets and bytes are tallied. This methodology of fingerprinting or determining a flow is scalable because a large amount of network information is condensed into a database of NetFlow information called the NetFlow cache. The collected data is stored in files of a limited period of 5-minutes to be used later for analysis.

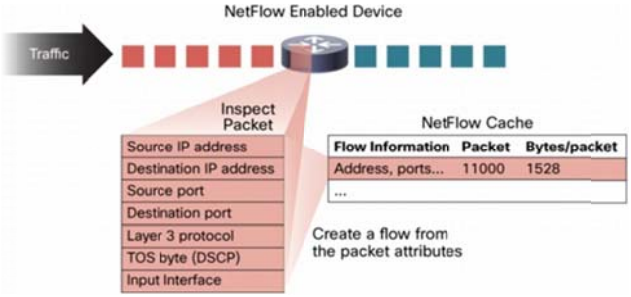


Figure 1. Creating a flow in the NetFlow cache[18]

Our study is based on CERNET backbone data, but the results could apply to general Internet traffic analysis. We will study the behaviors profiles depending on the available data; we are not going to study the behaviors of the whole Internet, because it's not possible to have the whole traffic of the Internet. The work is not limited to the managed domain, but it could be more general. The main focus is to be able to setup a model to study the behaviors of IP addresses.

IV. BACKGROUND

A. Entropy

We first introduce the concept of entropy, which is a measure of the uncertainty of a random variable. Let X be a discrete random variable with alphabet X and probability mass function $p(x) = P\{X = x\}$, $x \in X$. We denote the probability mass function by p(x), thus, p(x) and p(y) refer to two different random variables and are in fact different probability mass functions. The entropy H(X) of a discrete random variable X is defined by:

$$H(X) = -\sum_{x \in X} p(x) log(x)$$

Note that entropy is a functional of the distribution of X. It does not depend on the actual values taken by the random variable X, but only on the probabilities. Entropy (or Uncertainty) is a positive value $H(X) \geq 0$.

From the definition, when all observed values of the variable take the same value, which means no change in the results, the value of the Entropy is zero:

$$log\big(p(x)\big) = log(1) = 0$$

so

$$H(X) = 0$$

. On the other side if the observed values are totally different so the value of Entropy is the maximum value of the uncertainty:

$$p(x) = \frac{1}{N}$$

so

$$H(X) = H_{max}(X) = log(N)$$

B. Clustering

Clustering[19] is the unsupervised classification of patterns (observations, data items, or feature vectors) into

groups (clusters). Patterns within a valid cluster are more similar to each other than they are to a pattern belonging to a different cluster. Clustering is an unsupervised classification which is different from supervised classification in that there are no pre-classified (labeled) patterns. In the case of clustering, the problem is to group a given collection of unlabeled patterns into meaningful clusters. In a sense, labels are associated with clusters also, but these category labels are data driven; that is, they are obtained solely from the data. Data should be prepared for clustering by a sequence of processes like Feature selection, Feature extraction, and normalization as shown in Figure 2. Feature selection chooses distinguishing features from a set of candidates, while feature extraction utilizes some transformations to generate useful and novel features from the original ones. Both are very crucial to the effectiveness of clustering applications.

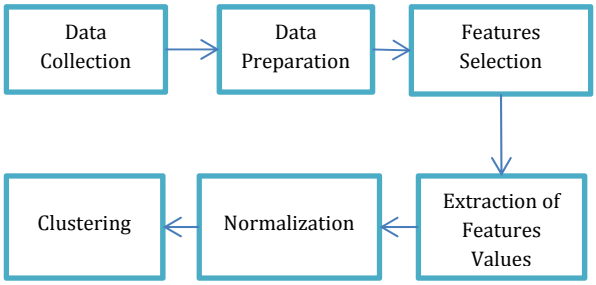


Figure 2. Major steps of host profiling procedure

DBSCAN[20] Clustering Algorithm: To find a cluster, DBSCA starts with an arbitrary point p and retrieves all points density-reachable from p. If p is a core point, this procedure yields a cluster, otherwise p is a border point (noise) and no points are density-reachable from p and DBSCAN visits the next point of the database.

V. 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.

Because it is not reasonable to setup a profile for each observed IP address, so they are classified. Classification or clustering of IP profiles will be based on their network traffic behavior to identify the service behind this IP address. 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 differentiate and 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 a very important step that needs to be specific to the problem. Currently, there is no study available for understanding and comparing the effect of feature selection in the context of NetFlow data. A combination of features will be used, some of them are directly extracted features, and others are calculated from the collected features using simple calculations or statistical analysis or obtained after applying techniques from the information theory like entropy (or Uncertainty). It's not possible to study all IP addresses or all clusters obtained, so the attention of this study will be focused on a few of the clusters or IP addresses which we call them the most significant.

VI. EXTRACTION OF THE MOST SIGNIFICANT IP ADDRESSES

It's not possible to monitor and profile every IP address appears over the internet, even each IP address in the trace, so we focus on the most significant IP addresses. The term "significant clusters of interest" were used in [4] by applying entropy based approach to cluster IP hosts on each dimension of the four-feature space, SrcIP, DstIP, SrcPrt, and DstPrt to extract the significant clusters of interest. The extracted SrcIP, DstIP clusters yield a set of "interesting" host behaviors (communication patterns), while the SrcPrt and DstPrt clusters yield a set of "interesting" service/port behaviors, reflecting the aggregate behaviors of individual hosts on the corresponding ports. In our research we depend on a more efficient and less cost method to extract the most active IP addresses that represent most of the flows in the trace. We have found that excluding 10% of the flows could means reducing the number of IP addresses that need to be analyzed in a very efficient way. In the following figure we can notice the number of significant clusters of interest from the total and distinct number of IP addresses, and because our study focuses on active flows initiated by the IP address, so we extracted the significant clusters of interest based on SrcIP. Let n denotes to the number of flows, m is the number of distinct elements of srcIPs, If X={x1, x2, ..,xm} is the complete list of source IPs, let p(xi) represents the possibility of appearance of xi in the flows of the trace during the period of study. We want to study IP behaviors over a long enough period to be able to get valuable profiles so that we need to extract the most significant clusters of interest. We select an epsilon value $\varepsilon = 0.1, 0.2$ to exclude the srcIPs that initiate flows less than 10%, 20% of the total flows, and analyze IP addresses that initiate more than 90% and 80%. The remaining significant srcIPs is the list of SrcIPs that initiate flows more than 90% of the total flows:

$$\sum p(x_i) > 1 - \varepsilon$$

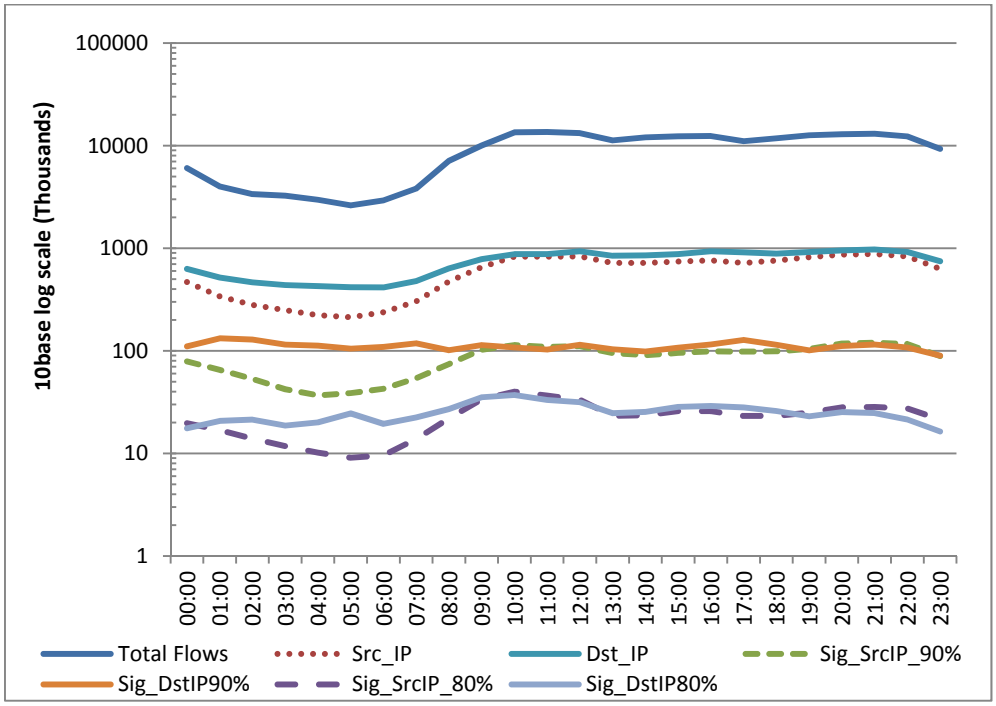


Figure 3. a logarithmic scale to base 10 to display the significant clusters of interest of Source and destination IP addresses that represent 90% and 80% of flows to the total number of IP addresses and the total number of flows over a complete one day with periods of one houre.

We developed an efficient algorithm to extract the list of significant SrcIPs which will be analyzed later. As we notice in Figure 3 that for periods of one hour, the maximum number of flows may reaches tens of millions with about one million of different source IP addresses. Analyzing this big number of IP addresses is impossible, so we select the most significant IP addresses, in the figure we may notice that if we exclude 10% of the flows we may get a list ten times less than the original of source IP addresses that initiate 90% of the total flows captured by netflow, and if we exclude 20% of the total flows we may get a list of 1/30 of the original distinct source IP addresses and this small list initiate more than 80% of the total flows. For our study, to get a more reliable and more reasonable results we have excluded 10% of flows and studied the 10% of source IPs that initiate more than 90% of the total flows.

VII. 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. We use only packet header information provided by NetFlow, we obtain direct and indirect features for each host. Direct features are retrieved directly without further computation, while Indirect features include those computed using multiple packets in a host's communication. Our focus will be on active communication carried by the profiled host and ignoring passive communication carried by other hosts. 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): the number of distinct IP addresses contacted by this host to which at least one packet is sent. This feature distinguishes the host community of peers that receive traffic from this IP. In other words the peers are the destination IPs to which at least one packet is sent to in the trace. This feature reflects the popularity of the IP node, and the importance of this feature comes from that this feature distinguishes one-to-one communications (like P2P or downloads) from one-to-several (like in web browsing) and one-to-many (like in netscans).
2. The ratio of the entropy of the first Destination IP byte to the entropy of the fourth Destination IP byte H(IP1)/H(IP4).
3. The ratio of the entropy of the second Destination IP byte to the entropy of the fourth Destination IP byte H(IP1)/H(IP4).
4. The ratio of the entropy of the third Destination IP byte to the entropy of the fourth Destination IP byte H(IP1)/H(IP4).

These features reflect the social or functional role of a host, these features will characterize the dispersion observed in the list of peers (or Destination IP addresses) associated with a Source IP. We need to study the distribution of IP addresses, but because IPs are not values to apply statistical measurements over the values of the IPs, also the complete distribution of the peers in the IP space would be too complicated to characterize and will not give the desired results, so we apply Shannon entropy S which measures the distribution dispersion. Entropy for IP distributions has been previously used in

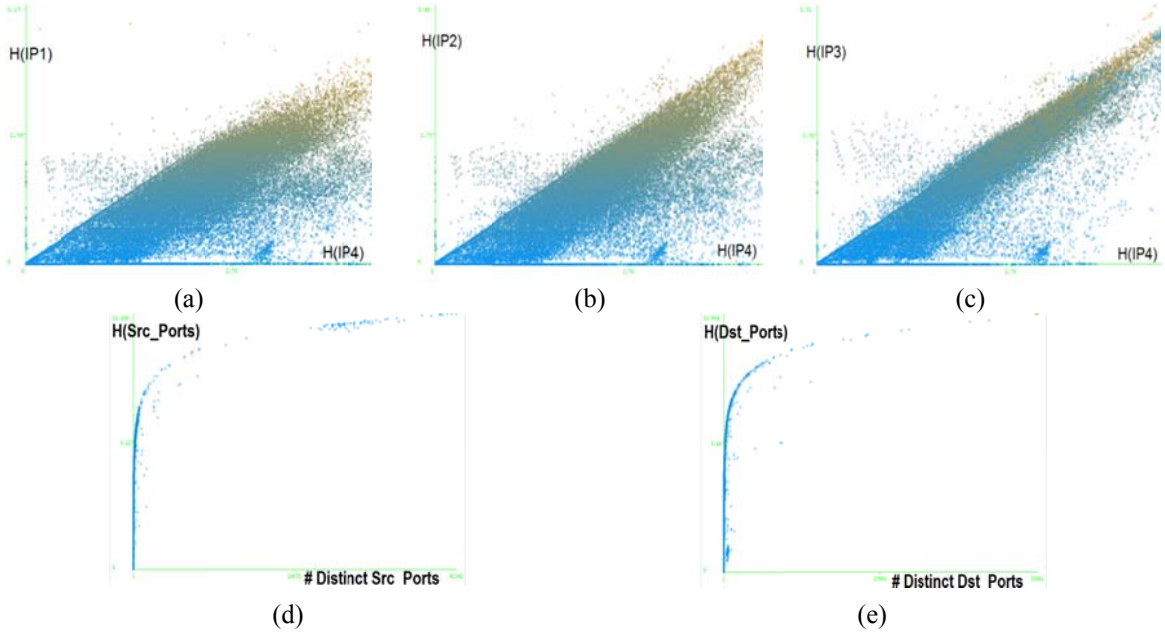


Figure 4. (a), (b), and (c) show the entropies of IP for the various hosts in traffic represented as a scatter plot of S(IP1), S(IP2) , S(IP3) vs. S(IP4). While (d), (E) show the Entropies of the Source and Destination Ports VS. Distinct source and destination ports

[10] where a ratio of entropies of the IP second and third bytes over the fourth byte where used, we found that adding the ratio of the first destination IP byte to the fourth byte will give a more accurate results in clustering IP hosts. 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. So the distributions of regular traffic inherit from this structure. 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. A large difference in these entropies is likely to betray scanning. Figure 4 shows the entropies of first, second, and third byte of IP addresses to the fourth byte of destination IP addresses for the various hosts in traffic represented as a scatter plot of S(IP1), S(IP2) , S(IP3) vs. S(IP4); each dot represents a host. Two different areas are apparent: S(IP1), S(IP2) , and S(IP3) << S(IP4).

5. The ratio of the number of source ports per the number of peers: this feature is very useful to reflect the role of the host or the node represented by the IP address, servers usually receive requests from clients on a single, 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. Large values of the number of source ports could means attacks like port scans.
6. The desperation of distribution in source ports: The number of distinct source ports itself may not reflect valuable meaning, so we use the desperation of distribution in source ports which will provide a valuable information of the communication pattern of the host, 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 so this port will be represented as one of these ports used by this host and will not reflect the frequency of using this port. So if this server sends 100 flows and 90% of them are http while the other flows belong to other different services, and another host sends 11 flows on different ports so the distinct numbers of ports are the same, while when applying Shannon entropy on source ports we get totally different numbers reflecting the distribution of the used source ports. So if the value of entropy is low, that means that one or some ports are used heavily on this host as source ports.
7. The ratio of the number of destination ports per the number of peers: as mentioned above, this feature is useful to reflect the role of the host or the node represented by the IP address. 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.
8. The desperation of distribution in destination ports: similar to feature number 6 also this feature reflects the distribution of destination ports. Figure 4 displays the Entropies of the Source and Destination Ports vs. Distinct source and destination ports. We can notice some strange values that represent non-equally distributed ports. As it's known for us that the entropy curve should take the logarithmic shape, but we may notice the points out of the logarithmic curve which represent IP nodes with special or anomaly behavior.
9. The mean number of packets per flow distinguishes elephant flows from mice flows (which are

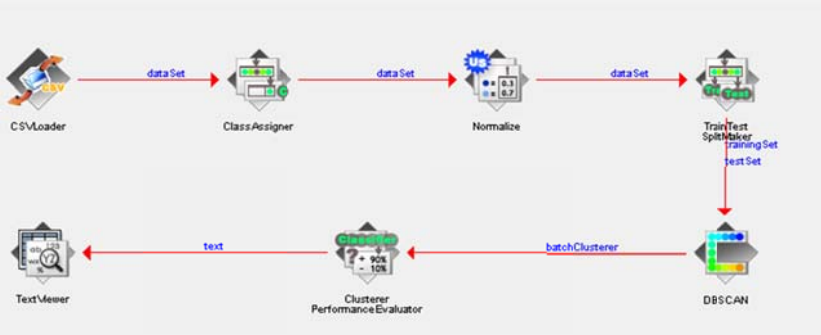


Figure 5. DBSCAN CLUSTERING USING WEKA

considered as non-connected flows and could be an attack).

10. The mean packet size reflects meaningful information in understanding the traffic produced where the small-size packets mostly consist of signaling traffic while large-size packets indicate data exchange.
11. The ratio of number of flows to the number of peers gives the mean number of flows to each destination IP which reflects consistency of traffic between these two hosts. While the flow is created by netflow during 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 the 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 differentiate between service providers that mostly use single protocols and normal clients that tend to be using a different protocols. IP protocol value comes from the flow record value (where 6=TCP, 17=UDP).
14. The entropy of type of application also differentiates between IP addresses representing nodes providing some services and clients that normally use different types of applications. (The type of application comes from the flow record where FTP=1, www=2, Mail=3, P2P=4, Service=5, Interactive=6, Mulitimedia=7, Voice=8, others=0).
15. Number of SYN_ACK sent by the host: To establish a connection, TCP uses a three-way handshake. Before a client attempts to connect with a server, the server must first bind to and listen at a port to open it up for connections: this is called a passive open. Once the passive open is established, a client may initiate an active open. The TCP three-way handshake in Transmission Control Protocol (TCP-handshake) is the method used by TCP set up a TCP/IP connection over an Internet Protocol based network.

SYN: The active open is performed by the client sending a SYN to the server. The client sets the segment's sequence number to a random value A.

SYN-ACK: In response, the server replies with a SYN-ACK. The acknowledgment number is set to one more than the received sequence number i.e. A+1, and the sequence number that the server chooses for the packet is another random number, B.

ACK: Finally, the client sends an ACK back to the server. The sequence number is set to the received acknowledgement value i.e. A+1, and the acknowledgement number is set to one more than the received sequence number i.e. B+1.

So When two computers attempting 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 its clients to start or end the session. So only service providers send this message therefor it's important to use the number of SYN_ACK messages as a feature of communication patterns of the hosts.

VIII. CLUSTERING AND RESULTS DISCUSSION

The above presented 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. We applied DBSCAN clustering algorithm using weka[21] as in the Figure 5 which includes the following main items:

- CSVloader: helps loading data from a csv file.
- Class Assigner: assign classes to the SrcIP
- Normalize
- Train Test Split Maker
- DBSCAN clustering
- Cluster Performance Evaluater
- Text Viewer to show the results.

After clustering, it's easily possible to notice some significant clusters like those presented in Table I:

A. Clients sending http 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

Table I.
SOME SELECTED CLUSTERS GENERATED BY DBSCAN

	Cluster label	1	2	6	19
	Number of Cluster Members	234	803	17	132
	Features	**Averages of the values of extracted Features of the cluster**			
1.	**Number of peers**	1	3	1549	2343
2.	**H_IP1/4**	0	0.072	0.271	0.364
3.	**H_IP2/4**	0	0.070	0.401	0.447
4.	**H_IP3/4**	0	0.083	0.995	0.805
5.	**Number of srcprts per peers**	50	0.891	0.001	0.0242
6.	**H_srcprt**	4	0.067	0	0.005
7.	**Number of dstprts per peers**	1	31	0.00084	6.38
8.	**H_dstprt**	0	4.117	0	8.34
9.	**Mean pkts per flow**	1.2	440	1	2
10.	**Mean pkt size (byte)**	590	1454	75	1225
11.	**Mean flows per peer**	56	44	1	8
12.	**Mean duration of flow (ms)**	6526	14695	0.0006	4559
13.	**H_prot**	0	0	0	0.0004
14.	**H_toa**	0.0027	0.008	0	0.0053
15.	**Number of SYN-ACKs**	0.0256	2.42	0	636

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.

B. P2P Traffic

Cluster with label 2 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 where: 1) all computers share equivalent responsibility for processing data. 2) Computers in a peer to peer network run the same networking protocols and software. 3) Peer to peer networks handles a very high volume of file sharing traffic by distributing the load across many computers.

C. Scanning a single port

As we notice in the values showed in Table I, the cluster labeled with number 6 the number of elements in this cluster is not big 17 SrcIPs, the traffic behavior of the hosts in this cluster is anomalous. 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 their transmission used only one protocol and a single type of application, and also we may notice that a single flow is made with the destination IPs. We may notice also that the changes in the third and fourth bytes of destination IPs 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.

D. Server traffic behavior

From the same table mentioned above we may notice the cluster with label 19 which includes 132 elements, they show a server traffic behavior based on their transmission, 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 which means a new port for each connection. And as it's known that clients request a service that is listening on a specific port on the server and assign a new (random) port number on the client, this new random port number is used as a destination port in the traffic transmitted from the server to the client. We can notice that the hosts in this cluster send a big number of SYN-ACK signals which is can't be transmitted 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. And as we have mentioned early when we selected features that servers tend to use a single protocol and one type of application more than others, we can notice that the value of entropy in the type of protocol and type of application is very low.

IX. CONCLUSION

The contribution of this paper includes: 1) discussion about the 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. 3) We analyzed IP nodes traffic behavior on a relatively long period of traces, which help to extract a more stable host's behavior. While previous studies focus only on host behavior for relatively short periods of 5 to 15 minutes, we extract host's behavior patterns over an hour which needs big data analysis to provide results in a reasonable time.

ACKNOWLEDGMENT

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 this work was sponsored by the National Grand Fundamental Research 973 program of China under Grant No. 2009CB320505, the National Nature Science Foundation of China under Grant No. 60973123, and the Technology Support Program (Industry) of Jiangsu under Grant No. BE2011173. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of those sponsors.

REFERENCES

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

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

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

[4] 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.

[5] 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.

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

[7] 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.

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

[9] 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.

[10] 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.

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

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

[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*.

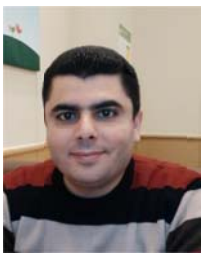
[18] "Introduction to Cisco IOS NetFlow - A Technical Overview," http://cisco.com/c/en/us/products/collateral/ios-nx-os-software/ios-netflow/prod_white_paper0900aecd80406232.html.

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

[20] 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.

[21] 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.

BIOGRAPHIES



Ahmad Jakalan, was born in Aleppo, Syria in 1982. Now he is a PhD candidate in the School of Computer Science and Engineering, Southeast University, China. His research interests are network security, network intrusion detection, and network traffic and host profiling. 2005 He has received his BS in Informatics Engineering from Aleppo University, Aleppo, Syria. 2011 He has received his MS in computer science and technology from Southeast University.

Jian Gong is a professor in the School of Computer 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.

Zhang Weiwei is a PhD candidate in the School of Computer Science and Engineering, Southeast University. His research interests are network security, network intrusion detection, and DNS traffic monitoring. 2007 He has received his BS in Software Engineering from Southeast University.

Qi Su is a Ph.D. candidate in School of Computer Science and Engineering, Southeast University, Nanjing, P. R. China. His research interests are network measurement and network management. He received B.S degree in computer science and technology from the Southeast University, Nanjing, P. R. China.