User Characterization through Network Flow Analysis

Vinupaul M.V
Centre for AI & Robotics
Bangalore,INDIA 560093
Email: vinupaul@cair.drdo.in

Raktim Bhattacharjee
Centre for AI & Robotics
Bangalore,INDIA 560093
Email: raktim@cair.drdo.in

Rajesh R.
Centre for AI & Robotics
Bangalore,INDIA 560093
Email: rajeshr@cair.drdo.in

G. Santhosh Kumar
Dept. of CS, CUSAT
Kochi,INDIA 682022
Email: san@cusat.ac.in

***Abstract*—One of the objectives of network security is to control the use of shared resources among users. In this regard, knowing the actual identity of network users is quite valuable to the intermediate nodes. The dynamic allocation of IP addresses and Network Address Translation(NAT) make it practically difficult, without any significant modifications to end user protocols and applications, to identify users by looking at network traffic. This work tries to establish network flow analysis as a viable method of user identification in such cases. We propose a supervised learning model that uses flow features to identify users within a given set. Based on our analysis of flow features, we introduce the concept of *flow-bundle-level features* which can be derived from the packet-level and flow-level features which are generally recorded by flow probes. We have identified a set of flow-bundle-level features which is able to identify users with a high degree of accuracy. This set comprises two types of features, user-features which are characteristic of the behaviour of an individual user and host-features which are the properties of the user's host platform. We also present an argument that, with the increasing penetration of personal mobile devices, a one-host one-user is going to be the norm and hence inclusion of host features will strengthen the user identification model. The model was validated, with four different supervised learning algorithms, using a data-set of flow records of 65 users. It was able to give a highest accuracy of 83%.**

Keywords: Network Flow Analysis, User Identification, Flow Features

I. Introduction

One of the objectives of network security is to control the use of shared resources among users. The ability to ascertain the identity of a network user is quite valuable for an intermediate network node like router as it opens up a new dimension of network provisioning. For instance, once the identity of a user is established, the network traffic of that user can be given different levels of treatment with respect to Quality of Service(QoS) or Quality of Experience(QoE). There are methods for user identification using custom protocols and applications. But these are practically difficult as they require significant changes to the large number of end host platforms and the intermediate networking nodes. We are looking at the problem of identifying users by looking at network traffic in the existing TCP/IP networking scenarios.

Though IP address is intended as a unique identifier in the Internet, the number of IPv4 addresses are limited and very small compared to the total number of hosts connecting to the Internet. As this limitation is bypassed by the use of techniques such as dynamic allocation of IP addresses and Network Address Translation(NAT), IP address cannot be used as a reliable method of user identification.

User identification based on behavioural patterns in user activities has already been proved as a successful model in works like analysis of keystroke patterns by Brown [1] and analysis of game-play activities by Kuan-Ta Chen [2]. This is an important area of study in the field of biometrics. We explore the scope for applying similar concepts in the context of the problem of identifying users from network traffic. The generation of network traffic flows depends, to a large extent, on the users' personal preferences, behavioural characteristics and psychological traits. Two simple examples of dependence of network traffic flows on users' behavioural traits are the set of favourite websites accessed and the average amount of time spent in favourite websites. These and similar idiopathic aspects generate distinguishing patterns of traffic flows, over a long period of time, that are capable of characterising different users. Supervised machine learning techniques can be used to train a classifier with known labeled data-set of users and then identify users from unknown traffic flows.

A flow is defined in RFC 5470 [3] as*"a set of IP packets passing an observation point in the network during a certain time interval, such that all packets belonging to a particular flow have a set of common properties"*. These common properties include packet header fields such as source IP, destination IP, source port, destination port and protocol. Flow-probes generally record certain configurable pieces of information corresponding to each flow in the network traffic, identified as a flow record. A flow record includes: i) packet-level features - those available directly as fields in the headers of each packet of the flow and ii) flow-level features - those that need to be computed based on values collected from all the packets in a flow. Examples of packet-level features are source/destination IP address and port and protocol type. Examples of flow-level features are like average packet size, flow data rate and flow duration.

In the proposed user characterization system we introduce and use flow-bundle-level features. Flow-bundle-level features represent a bundle of flows and are computed from the values of flow-level features of all flows in the bundle. Examples

978-1-5090-1281-7/16/$31.00

of flow-bundle-level features are average inter-flow-gap, flow rate (number of flows per hour), average flow duration. The flow-bundle-level features are an important concept because individual flows in isolation are not capable of characterising the user. When grouped together into bundles, flows can reveal certain distinguishing patterns characteristic of users' behaviour. Often these patterns are recurring and unique to an individual.

The advancements in Software Defined Networking (SDN) and its wide adoption make this work all the more relevant and important. As against the rigid and non-programmable traditional hardware routers, the flexible and programmable SDN framework is conducive to the experimentation and adoption of this concept. A SDN switch can be programmed to export flow information to its controller which in turn can attempt user identification and then insert separate rules into the switch based on the discovered user identity.

In this work, we try to establish network flow analysis as a viable method of user identification. It can be considered as a fusion of two major areas, network flow analysis and user identification based on behavioural activities. And it uses machine learning techniques. It is one of the few of its kind that attempts to use network flow analysis to recognise the behavioural patterns and thereby identify users. This work takes the assumption that there is a strong coupling between a user and the host. This is especially true with an increasing penetration of personal mobile devices like smart phones, tablets and laptops. Each host has certain unique features determined by its make and operating system installed. The user features when combined with host features can identify a user with high accuracy level. The survey report [4] which shows that, by 2019, 71% of the internet traffic will be from mobile gadgets further reinforces our assumption.

This paper proposes the use of machine learning techniques to identify users from network traffic flows. It also covers an analysis of flow features and different taxonomies to classify them, introduces the concept of flow-bundle-level features and presents a set of features that can identify users with high accuracy. We implemented the concept by building a classifier using the Scikit learn[5] library. The model was validated with a data-set of 65 users and achieved an accuracy of 83%.

The rest of the paper is organised as follow. Section 2 describes the related work in this area, Section 3 explains the rationale of selecting different flow features for this work, Section 4 describes the model selection and test result. Section 5 describes future work and finally, Section 6 concludes the paper.

II. Related Works

As noted above this work is a fusion of two major areas, network flow analysis and user identification based on behavioural activities. Of late myriads of research activities have been taking place in the field of network flow analysis. Network flow analysis can be perceived from three perspective namely Administration, Law Enforcement and Network Security. Administration perspective better equips you in troubleshooting, resource planning and forecasting, law enforcement perspective helps you discover useful intelligence, and Network Security perspective enables you to detect policy violations and security incidents. The survey paper by Bingdong Li et al. [6] gives an elaborate description about data-sets, perspectives, methodologies, challenges and future directions of network flow analysis. One of the major applications of network flow analysis is application classification. Works of Andrew W. Moore et al. [7], M Crotti [8], S Zander [9] and L Bernaille [10] clearly prove that flow analysis with machine learning can provide high level of accuracy in application classification. Flow analysis is also used in the field of intrusion detection. Celeda, Francois and Ren et al uses network flows analysis in identifying security incidents like botnets and worms.

Identifying users based on behavioral activity is not a new concept. One of the prominent works is that by Brown [1] which identifies users by analysing keystroke patterns with neural network. A model is created for each user's typing style which is later used for identification. Yun [11] uses walking pattern of an individual as parameter for his activity based user identification system. His system can recognise the registered users at the rate of 92%. Ikehara, C.S [12] used the force applied by an user to a computer mouse as a signature for user identification. Again, Kuan-Ta Chen [2] proposes a new biometric for human identification based on users' game-play activities with an accuracy of 90% with 20 minutes detection time.

The concept of identifying users based on network flows is still quite rudimentary. This was first introduced in the paper by Nikolay Melnikov et al. [13] which demonstrates that length of flow is a distinguishing feature of each individual. There is no continuation of this work further.

III. Proposed model for user identification

We propose a supervised learning model that uses features extracted from flow records to identify users within a given set. This model helps establish network flow analysis as a viable method of user identification in the intermediate nodes of TCP/IP networks.

A. Analysis of Flow features

There are a large number of features associated with network flows which could potentially be used by the the supervised learning model for user identification. In order to identify a small number of candidate features from among these, to be used in the proposed model, we did a detailed analysis of flow features. In the process of analysis, we looked at the various ways in which flow features can be classified. These different taxonomies help in a proper and efficient collection of features from network traffic, a better understanding of the relevance of features and thereby selection of candidate features.

In the first taxonomy, flow features are divided into two types - basic flow-record features and derived features. Basic flow-record features are those available as fields in the flow records generated by standard flow probes. These flow records

conform to the widely adopted formats like Netflow and IPFIX(RFC 7012) [14] and can include any of the more than 400 listed fields. Derived features are those which are generally not available as explicit fields in the flow records generated by flow probes.

In the second taxonomy, flow features are divided, based on from where and how the feature is collected, into three types - packet-level features, flow-level features and flow-bundle-level features. Packet-level features are those available directly as fields in the headers of each packet of the flow and can be directly extracted from any of the packets. Examples of Packet-level features are Source/Destination IP address and Port, protocol type. Flow-level features are those that need to be computed based on values collected from all the packets in a flow. Examples of flow-level features are statistical parameters like average packet size, median of packet size, flow data rate, flow duration, average packet interval. Flow-bundle-level features are those that need to be computed based on the values of flow-level features. Examples of flow-bundle-level features are average inter-flow-gap, flow rate(number of flows per hour), average flow duration. Flow-bundle-level are not generally available as part of flow records and need to be derived from flow records.

In the third taxonomy, flow features are divided into two types - user features and host features. User features are characteristic of an individual user. It is determined by the behavioural and psychological traits of the user. Host features are dependent on the host platform used by the user. These features are determined by the operating system, the kernel version, the applications installed and so forth. The host features are not generally controlled or influenced by the user.

B. Feature collection from flows

The basic flow-record features can be collected using a standard flow probe which conforms to the netflow format. The flow probe can be configured to record only the selected features. The derived features can be obtained either by using a custom flow probe or from the netflow records by using appropriate computations.

C. Generation of derived features

For the purpose of generating derived features, flows are grouped together based on the host IP addresses in the network. The features we derive are the flow-bundle-level features representing a group of N flows. These flow-bundle-level features (f_i) are derived from the flow-level features of the flows in a group. The values of derived features constitute a sample for the learning model. We have used the concept of sliding window for the purpose of bundling the flows of a host. Suppose the flow records of a host in the data-set $1, 2, 3, ...n$ are represented on a number line and let W be conceptualised as a window which can cover N (such that $N < n$) flows when superimposed over the line. The N flows falling inside the window frame together generates one sample ($S_i = \{f_{i1}, f_{i2}, ...f_{im}\}$). The window W starts from the left end of the line covering the N flows $(1, 2, ..N)$, then gradually moves forwards subsequently covering $(2, 3, ..N + 1), (3, 4, ...N+2)$.... In the process of sliding, the sets of flows falling inside window frame in subsequent positions generate the set of samples $S_1, S_2, S_3...S_n$ for that host.

D. Feature selection

We are basing the selection of features on the analysis and classification of features given above. The selection is based on heuristics and domain knowledge. These features are selected in such a way that for each user it remains fairly consistent over long periods of time, there are significant variation among features among different users and there are strong similarity among features of same user.

With respect to the basic flow-record features and derived features, we have selected only certain features from the derived category and these are the flow-bundle-level features. The flow-bundle-level features are able to reveal the patterns in flows better than packet-level and flow-level features. It is the values of derived feature set that constitute a sample for the proposed learning model, rather than a normal flow record.

In selecting features for the proposed model, with regard to the user features and host features, we have drawn features from both categories. We assume that there is a strong coupling between an user and a host system. One host system is used by only one user. This is especially true with an increasing penetration of personal mobile devices like smart phones, tablets and laptops. With this assumption we have identified a set of features, for the user identification problem, comprising almost equal number from both the classes. We have observed in our experiments also that host features strengthen the user identification model.

The list of selected features and the relevance of each are given below:

1) User features:

- **Destination IP dispersion:** It is the number of unique destination IP's visited in a window frame. Its value depends on the user's interest and psychological traits and varies widely among different users. Data analysis shows that standard deviation of IP Dispersion over time for the same user is quite less compare to that for different users. The Destination IP Dispersion for two randomly selected users is plotted in Fig.1 and clearly shows a wide variation between the two users.
- **Site of interest:** It is the destination IP contacted by an user the maximum number of times in a group of N flows. This feature depicts the site most frequently visited by the user. Analysis of the data-set shows that, site of interest has a good variability among different users.
- **Destination IP cohesion:** This feature indicates the size of the biggest cluster in the set of destination IPs and is a measure of how frequently the user is visiting that site of interest. Its value is the number of times the site of interest, the most frequently visited destination IP, is visited in a window frame.
- **Application diversity:** It quantifies the number of different application layer services used by an user. It is

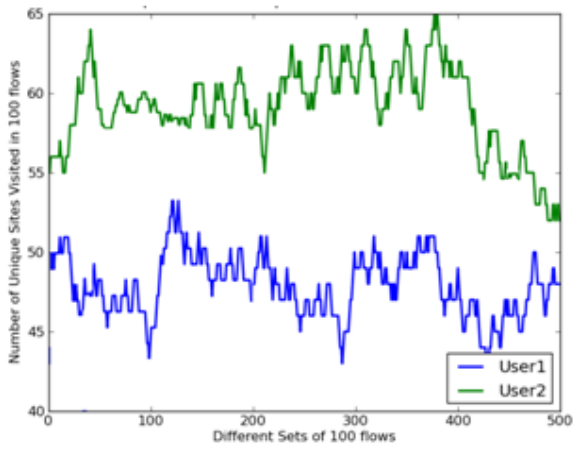

Fig. 1. Destination IP Dispersion for 2 random users

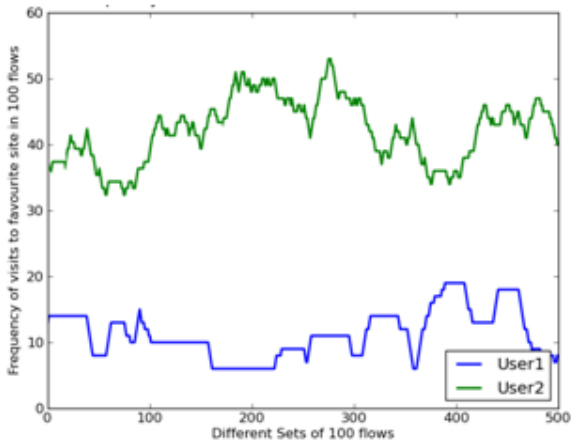

Fig. 2. Number of times the favourite site is visited

measured by the number of unique destination ports in a group of N flows. Destination ports in the flows represent the services accessed by an user and the unique destination ports represent how many different services the user is accessing.

- **Favourite application:** This feature specifies the application layer service that a user is most interested in. It is the top destination port among the group of N flows. This feature becomes significant in the case of mobile devices where different applications use different port numbers. There are some users more fond of using Hike messenger where as there are some others more fond of Whatsapp. It is this natural difference among users with respect to the favourite application that this feature tries to capture.
- **Browsing duration:** This feature tries to profile the average amount of time a user spends in surfing websites. It is measured as the median of flow durations in a group of N flows. It also gives an approximate idea about the uploading/downloading activities. Different users use the Internet for different purposes. One user might use it mostly for mail exchanges while another might use it for multimedia download. User's purpose has a direct implication on the flow duration.

2) Host features:

- **Median of source ports:** It is the median of the values of source ports of all the flows in the bundle. When a host system is acting as a client the source ports are drawn from the ephemeral port range by the operating system kernel as per RFC 6056 [15]. This ephemeral port range varies with operating system and kernel and hence hosts with different OS and kernel show a variation in this value. Median of source ports vaguely captures the type of operating system and kernel version in the host.
- **Number of unique source ports:** When a host is acting as a client and establishes a connection with a server, a random source port is selected for the connection based on RFC 6056. Re-use of the same port for subsequent connections is purely kernel dependent. When a host system is behaving as a peer or as a server a pre-designated fixed port is used as the source port and is re-used again and again. Thus the number of unique source ports in a bundle of flows varies widely depending on the OS kernel and the services hosted.
- **Number of unique TTLs:** TTL(Time To Live) in an IP packet indicates the maximum time the packet is allowed to remain in the network. It is specified as the maximum number of hops the packet can travel. The TTL field in a packet is initialized by the OS kernel and then decremented by the network nodes at each hop. Different OSes are using different values for this initialization and hence the TTL value in packets is generally used in distinguishing and fingerprinting the host OSes. The presence of different TTL values in different packets from the same host could indicate that the host system is having multiple OSes with hyper-visor.
- **Top TTL:** As noted earlier, the initial TTL value is a distinguishable characteristic of OSes. If a host system has multiple OSes, multiple TTL values can be observed in the flows. Hence the Top TTL gives an indication of the only OS or the most prominent OS in the host system.

These eight features have been empirically found to be good enough to characterise users using the supervised learning model. This was validated using a data-set of 65 users. The set of experiments that were and the results are described in the next section.

E. Classification

We have defined the problem as identifying users within a given fixed set using the traffic flow features. It is a supervised learning problem, more specifically a multi-class classification problem where the classes are the various user identities. Classification is defined as mapping each data item in one of the several classes. Classification algorithms work in two phases, the training phase wherein a model is built from the training data - feature vector and label - and the prediction phase wherein the model uses a feature vector as input to predict the label.

In our context the classifier is trained on a set of flows, collected from a fixed set of users, which are labled with the user identities. In the classification phase, an unlabled set of flows, again collected from the same set of users, is given as input to the classifier and the classifier predicts the user identity label for each of the flows.

We used four different classification algorithms in our experiments, k-Nearest Neighbour (for k=3 and k=7), Random Forest and C4.5 Decision Tree. The results of each and the comparison are given in the next section.

F. Performance Evaluation

We evaluated and compared classifier performance using the various metrics derived from the four parameters -True Positive(TP), False Positive(FP), True Negative(TN) and False Negative(FN)- described in the standard confusion matrix. The metrics used are the following:

- Receiver Operating Characteristics (ROC) graphs: It plots TP rate versus FP rate for every possible classification threshold. The ROC curves are compared by calculating the Area Under the ROC curve(AUC) whose value ranges between 0 and 1. The higher the value of AUC the better is the classifier. ROC quantifies classifier ability to avoid false classification.
- Accuracy:It is the one which focuses on overall effectiveness of the classifier.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

- Precision(P):It gives a good estimate of what fraction of the prediction made by the classifier make is really true.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

- Recall(R):It estimates the classifier's ability to identify the percentage of true results out of all true results.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

- F1 Score: It is a metric derived out of Precision and Recall. F1 Score is calculated as:.

$$F1Score = \frac{2 * Recall * Precision}{Recall + Precision} \tag{4}$$

IV. Experiments and Discussion

A. Data-set

We validated our concept using a data-set of flow records of 65 users. The labeled data-set was collected from an university network wherein a group of 65 volunteers were asked to log their personal flow traces for a period of 7 days on their system. So as to maintain the privacy of each individual, only the first 64 bytes of each packets were logged. The data were recorded in pcap format. The volunteers were instructed not to allow their systems to be used by other users during the data collection. This ensured that one system is used by only one user and the flow traces collected a system is characteristic of the one user.

B. Feature extraction

The desired features were extracted from this data-set in two steps. First the basic flow-record features were extracted from the data-set and then the derived features at the flow-bundle-level were generated from the flow-records. The extraction of flow-records from the pcap file was done using a flow-probe, developed in-house by our team, which takes pcap file as input, constructs flows and stores the flow records in mysql database. Each flow-record, has these fields: source IP, destination IP, source port, destination port, protocol, flow direction and time duration. We analysed the flow-records and found that majority were TCP sessions and there were a only small percentage of UDP and ICMP. We filtered out and used only the TCP flows for our experiments. There were total 12,48542 TCP flows, ranging from 15,435 to 35478 per user. Then the derived features were generated from these flow-records using the concept of sliding window with $N = 100$. The generated flow-bundle-level features constituted samples for the classifier.

C. Classification and validation

We experimented with classifying the flow-bundle-level samples as belonging to the various users in the set. The classification was attempted with the four classification algorithms mentioned in the previous section and the results compared. The implementation was done using the scikit-learn machine learning library in Python. The classifier models in each case were cross-validated using 10 fold stratified cross validation.

Classifier	Accuracy	Precision	Recall	F1-score
Dec Tree	0.61	0.74	0.73	0.73
KNN7	0.67	0.76	0.67	0.71
KNN3	0.69	0.77	0.69	0.72
rand Forest	0.73	0.86	0.75	0.78

TABLE I
COMPARISON OF CLASSIFIERS WITH ONLY USER FEATURES

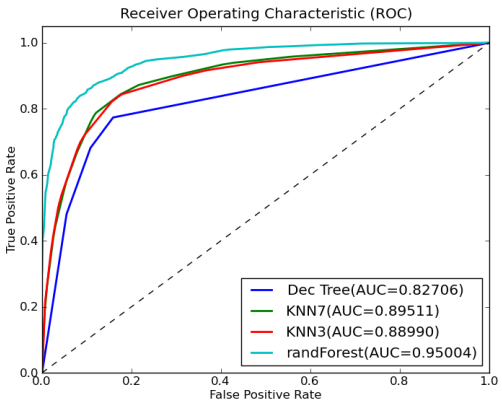

Fig. 3. ROC curves of classifiers with only user features

With regard to the performance of classifiers, we carried out two types of comparisons in our experiments. One is the comparison of different classification algorithms. The other is the comparison of feature sets with and without host features.

The comparisons were done in terms of accuracy, precision, recall, F1-scores and ROC Curves.

Table 1 shows the comparison of different algorithms when host features are not used and Table 2 shows the comparison of different algorithms when host features are used. Fig.5 and Fig.6 shows the corresponding ROC curves for the two cases. It can be observed that Random Forest outperforms other algorithms in all terms and that the host features significantly improve the classifier performance for any algorithm.

Classifier	Accuracy	Precision	Recall	F1-score
Dec Tree	0.70	0.80	0.82	0.81
KNN(7)	0.78	0.83	0.78	0.80
KNN(3)	0.78	0.83	0.79	0.80
rand Forest	0.83	0.91	0.85	0.87

TABLE II
COMPARISON OF CLASSIFIERS WITH BOTH USER AND HOST FEATURES

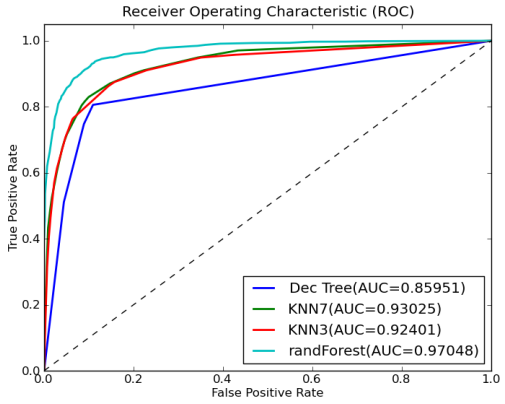

Fig. 4. ROC curves of classifiers with both user and host features

V. Conclusion and Future Work

In this paper, we have shown that network flow analysis is a viable method for user identification. Based on our analysis of flow features, we introduced the concept of flow-bundle-level features and identified a set of flow-bundle-level features which can identify users within a given set. We conducted experiments using the proposed supervised learning model which was found to give a good accuracy of 83%. The results show that the classification performance improves when host features are included. One of the challenges in using the flow data for user identification is the dynamic and evolving nature of the Internet protocols and applications. The performance of the proposed model need to be further validated over a long period of time by deploying it in a real user context. We can also explore the performance of other potential features which can be extracted from the flows.

VI. Acknowledgment

The authors would like to thank Director CAIR for all the support. We would also like to thank Shri. A.V. Sahadevan for extending the support for setting up the lab and guiding us through the experiments and providing comments on the initial draft of this paper.

References

[1] Brown, Marcus, and Samuel Joe Rogers. "User identification via keystroke characteristics of typed names using neural networks." International Journal of Man-Machine Studies 39, no. 6 (1993): 999-1014.
[2] Chen, Kuan-Ta, and Li-Wen Hong. "User identification based on game-play activity patterns", Proceedings of the 6th ACM SIGCOMM workshop on Network and system support for games, pp.7-12. ACM, 2007.
[3] Architecture for IP Flow Information Export 2009. https://tools.ietf.org/html/rfc5470
[4] Internet Society Global Internet Report 2015, IS_web.pdf. http://www.internetsociety.org/globalinternetreport/assets/download/
[5] Scikit Learn library. http://scikit-learn.org.
[6] Li, Bingdong, Jeff Springer, George Bebis, and Mehmet Hadi Gunes. "A survey of network flow applications." Journal of Network and Computer Applications 36, no. 2 (2013): 567-581.
[7] Moore, Andrew W., and Denis Zuev. "Internet traffic classification using bayesian analysis techniques." In ACM SIGMETRICS Performance Evaluation Review, vol. 33, no. 1, pp. 50-60. ACM, 2005.
[8] Crotti, Manuel, Maurizio Dusi, Francesco Gringoli, and Luca Salgarelli. "Traffic classification through simple statistical fingerprinting." ACM SIGCOMM Computer Communication Review 37, no. 1 (2007): 5-16.
[9] Zander, Sebastian, Thuy Nguyen, and Grenville Armitage. "Automated traffic classification and application identification using machine learning." In Local Computer Networks, 2005. 30th Anniversary. The IEEE Conference on, pp. 250-257. IEEE, 2005.
[10] Bernaille, Laurent, Renata Teixeira, Ismael Akodkenou, Augustin Soule, and Kave Salamatian. "Traffic classification on the fly." ACM SIGCOMM Computer Communication Review 36, no. 2 (2006): 23-26.
[11] Yun, Jae-Seok, Seung-Hum Lee, Woon-Tack Woo, and Je-Ha Ryu. "The user identification system using walking pattern over the ubifloor." In Proceedings of International Conference on Control, Automation, and Systems, vol. 1046, p. 1050. 2003.
[12] Ikehara, Curtis S., and Martha E. Crosby. "User identification based on the analysis of the forces applied by a user to a computer mouse." In System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference on, pp. 7-pp. IEEE, 2003.
[13] Melnikov, Nikolay, and Jrgen Schnwlder. "Cybermetrics:User identification through network flow analysis.", Mechanisms for Autonomous Management of Networks and Services, pp.167-170. Springer Berlin Heidelberg, 2010.
[14] Information Model for IP Flow Information Export (IPFIX) 2013. https://tools.ietf.org/html/rfc7012
[15] Recommendations for Transport-Protocol Port Randomization. 2011. https://tools.ietf.org/html/rfc6056