## **Refid: 2, An Empirical Investigation of Filter Attribute Selection Techniques for High-Speed Network Traffic Flow Classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Future research may involve conducting more experiments, including other filter attribute selection techniques and more datasets from other traffic flow |  | the classifier model built on the smaller subset of attributes has a comparable (no significant difference) performance to that built with a complete set of attributes. This would benefit the metrics collection, model calibration, model validation, and model evaluation times of future traffic flow classification |

## **Refid: 8, Dynamic Network Traffic Data Classification for Intrusion Detection Using Genetic Algorithm**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  | Moreover, the proposed classifier shows impressive performance while detecting normal behavioral traffic data patterns compared to other efficient existing classifiers as discussed in the previous section. Therefore, our proposed classifier can efficiently meet the needs for designing an up-to-date intrusion detection system with providing solutions to the limitations of existing ones operating in modern dynamic network environment. |

## **Refid: 9, Feature Selection in the Corrected KDD-dataset**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| further investigations are required to be carried out with real data for future work. |  | The proposed features were tested on the NSL-KDD dataset and the results demonstrated that the proposed features produce high detection rates |

## **Refid: 12, Data Summarization for Network Traffic Monitoring**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| More focus should be given to real-time summarization, so we will direct our research efforts into stream data summarization methods, and distributed summarization. |  | The running times of classifiers (or any other data-mining algorithms) using summarized data can be a fraction of the running time if the original dataset is used, while approximating similar results. This can save time in critical applications, where data analysis is the bottleneck of operations. |

## **Refid: 13, Hybrid Classifier Systems for Intrusion Detection**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## **Refid: 18, Decentralized Detection of Network Attacks Through P2P Data Clustering of SNMP Data**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In our simulations we explored just a little part of possible configurations: experiments could be made on further network topologies and for different distributions of data. | Compared to the results of the simulations, the final system may present even better performances with some improvements on the mining algorithm or using an approach different from k-means (which is relatively trivial) |  |

## **Refid: 24, Payload Modeling for Network Intrusion Detection Systems**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Future work in the areas of noise-removal and false positive mitigation can also prove valuable to payload-based approaches to IDS’s | noise-removal and false positive mitigation | One of the challenges faced by anomalybased systems is the availability of good, attack-free training data. Since training data collected under uncontrolled circumstances carry no guarantee of being attack-free, using such data to train can make an IDS susceptible to training based attacks. |

## **Refid: 32, Towards the Reduction of Data Used for the Classification of Network Flows**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Future works should concentrate on the reduction of false positive errors and the reduction of computational cost of the data transformation by eliminating the input attributes having a minor impact on the reduced features. Moreover, other dimensionality reduction techniques are planned to be included in the proposed framework |  | Taking into account the excessive volume of data that has to be constantly analysed for possible misuse of the internet services, extensive experiments aiming at selecting the appropriate reduction of the data are fully justified. |

## **Refid: 47, An Intrusion Detection System Using Network Traffic Profiling and Online Sequential Extreme Learning Machine**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| The proposed technique addresses various issues of IDS and network traffic dataset. Beta profiling with large threshold distance may affect the performance of IDS due to inclusion of biasness. In future, biasness reduction model may be proposed. This model may provide an opportunity to reduce the training dataset at par memory requirement. Some other feature selection technique like multi-objective optimization may be tested with proposed technique. The effectiveness of alpha and beta profiling can also be tested on other than network traffic datasets. |  | The network traffic dataset is huge and imbalanced. Distributions of connections for some protocols are higher than others. Intrusion detection system (IDS) with memory and time constrains find it difficult to process whole dataset. IDS also suffer low accuracy and high detection rate. I |

## **Refid: 48, A Multi-step Outlier-based Anomaly Detection Approach to Network-wide Traffic**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Our future work includes developing a parameter free hybrid outlier detection approach that combines both the distance and density based techniques for mixed type data to efficiently detect more types of attacks in network-wide traffic | Our proposed approach has two limitations: (i) It is dependent on proper tuning of a parameter called τ with respect to a dataset using a heuristic method, (ii) It does not work directly on categorical and mixed types data | Lowering the percentage of false alarms |

## **Refid: 50, Characterizing Network Traffic by Means of the NetMine Framework**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## **Refid: 51, An In-depth Analysis on Traffic Flooding Attacks Detection and System Using Data Mining Techniques**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Our future work will be made on a more concrete research that can support the establishment of policies for intrusion response systems and intrusion protection systems. Also, our ongoing works include addressing the presence of things on the Web and further validate the ideas of security in various Web environments. |  |  |

## **Refid: 52, A Framework for Application-driven Classification of Data Streams**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| To improve efficiency of the TS3 VM model, we will study how to train linear TS3 VM model in linear time for high dimensional sparse data. We will also attempt to improve accuracy of the TS3 VM model by incorporating kernel functions. |  | : concept drifting, large volumes, and partial labeling |

## **Refid: 53, History Guided Low-Cost Change Detection in Streams**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| The idea of deploying low-cost s-monitors into the data space can be applied to various applications. People often think of sampling when large amounts of data need to be processed, similarly, s-monitors can help in many occasions when expensive models are needed for stream change detection. We intend to apply our approach to other types of models and to look for good s-monitors with theoretically guaranteed accuracy. When history analysis is used to guide the s-monitor selection, the gradual removal of stale knowledge can be necessary for some applications. How to efficiently downgrade the effect of stale data is also an interesting problem |  |  |

## 

## **Refid: 56, Detecting Anomalies from Big Network Traffic Data Using an Adaptive Detection Approach**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## **Refid: 58, RFAODE: A Novel Ensemble Intrusion Detection System**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| we will apply evolutionary approaches for IDS to classify network traffic data on various data sets. |  |  |

## 

## **Refid: 59, A Data Mining Framework for Securing 3G Core Network from GTP Fuzzing Attacks**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| The future work also includes a thorough analysis of the processing overheads of the proposed framework to make it deployable in a real environment. |  |  |

## 

## **Refid: 63, Local Outlier Factor and Stronger One Class Classifier Based Hierarchical Model for Detection of Attacks in Network Intrusion**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  | searching for deviation to locate an attack in real time |

## 

## **Refid: 64, Intrusion Detection with Evolutionary Learning Classifier Systems**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| we would explore the use of adaptive mutation in GA search (see Hurst and Bull 2003), an adaptive covering interval mechanism, and alternative representations. | In these experiments we did not explicitly consider the processing times of these systems, another crucial aspect of intrusion detection. In future, we intend to implement these systems in a parallel and distributed framework to improve their scalability and speed | genetic search with this representation faces serious challenges in the type of high-dimensional real-valued space encountered here. Our optimisation of the covering interval helped somewhat but a more general and effective solution is needed |

## 

## **Refid: 69, Peer-to-Peer Traffic Identification by Mining IP Layer Data Streams Using Concept-Adapting Very Fast Decision Tree**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| in the future as more and more P2P applications use dynamic ports. So investigating the techniques which enhance the accuracy of P2P traffic classification by exploiting the unlabeled examples is our next work. |  |  |

## 

## **Refid: 74, Online Incremental Learning for High Bandwidth Network Traffic Classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In future, we will implement this work in reprogrammable hardware so that it can perform inline classification of live network traffic.. |  |  |

## 

## **Refid: 109, An Integrated Model for Prediction of Loading Packets in Network Traffic**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In future we shall examine if ambiguity in clustering affects the performance of the proposed model. |  |  |

## 

## **Refid: 133, Sub-space Clustering, Inter-clustering Results Association & Anomaly Correlation for Unsupervised Network Anomaly**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 134, Relational Network-service Clustering Analysis with Set Evidences**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In this work, only connectivity information is considered for clustering. In the result from real data, we do find 1 to 2 big clusters that contain many ports and applications. They are the results of some “universal” applications that employ many services (like explorer.exe). Other relations like the interactions between applications and users can be used in the future work to eliminate the influence of the “universal” applications. Since the network traffic is generated by users day after day, it is also desired to have an online version of the clustering algorithm. The main difficulty for an online algorithm is that relational clustering results usually depend on the order of adding new records |  |  |

## 

## **Refid: 149, Automatic Discovery of Botnet Communities on Large-scale Communication Networks**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| we will evaluate our approach on the P2P community and measure its performance on P2P based botnets. Until the deadline of the paper we have not received any P2P botnet traffic from the honeypot and we also attempted to search the source code of some well-known P2P bots (e.g. Rustock, Nugache and Peacomm) from the public malware sharing website so that we can run it and collect P2P botnet traffic traces on our testbed network (fortunately we will get the storm P2P .pcap data from the German Honeynet Chapter [46]). Also some novel P2P botnets construction methods have been proposed and investigated in [44, 45], and in summary, we will focus on the detection of existing and new appeared P2P botnets in the near future. | one is legal issues related to privacy and the other is that it is impossible to identify encrypted traffic | How to identify applications for network traffic?  How to detect new (or recent) appeared botnets? |

## 

## 

## **Refid: 166, An Autonomic Traffic Analysis Proposal Using Machine Learning Techniques**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| As future works, we propose to detail the autonomic architecture with a more fine selection of the pair features-model, that will integrate metrics from the Traffic analyzer and QoS manager. Additionally, more implementation matters will be treated such as on-line tests, complexity, resource management, time response, etc |  |  |

## 

## **Refid: 167, A Supervised Machine Learning Approach to Classify Host Roles on Line Using sFlow**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| As a future work, we will apply the knowledge we obtained in these experiments to build a network security analysis and anomaly detection system. We will also classify more granular roles of hosts in order to build signatures for each host of the network, and investigate more features and optimize features by feature selection algorithms. |  |  |

## 

## **Refid: 168, Machine Learning for Encrypted Malware Traffic Classification: Accounting for Noisy Labels and Non-Stationarity**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  | the scale of the data, demand for very low false positive rates, evolving data streams, and noisy class labels.  the network security domain |

## 

## **Refid: 169, Causality-based Sensemaking of Network Traffic for Android Application Security**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| For the future work, we plan to deploy our tool to collect Android network traffic and detect more malicious apps via crowdsourcing. Also, we plan to extend our solution for real-time triggering relation inference and online detection. | Our dependence analysis model aims at analyzing the network activities of Android devices, while it has certain limitations and possible evasions | Lack of referrer  Diverse network traffic from apps  Automatic notifications and updates |

## 

## **Refid: 171, A Preliminary Investigation on the Identification of Peer to Peer Network Applications**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Future work should investigate optimization of P2P application identification results by using another algorithms such as SOM, Fuzzy C-Means, improving current algorithms or developing a new algorithm. In addition, the FPR was high for the training and testing for UTorrentDS and BitTorrentDS data sets for most the DM algorithms. Generating a bigger dataset and implementing these algorithms over this set may help us overcome these drawbacks. Finally, other traffic analysis approaches can be applied to see their effectiveness on this task. |  |  |

## 

## **Refid: 173, BotFinder: Finding Bots in Network Traffic Without Deep Packet Inspection**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| The sandboxed training environment can be improved to better circumvent malware authors to probe for virtual machine settings and react by stopping all activity. 2: A separate problem – of the general anti malware research community – is the classification of malware to families. Currently we rely on Anubis signatures and VirusTotal labels which yield sufficiently good results – especially because we drop small clusters in the model building process and thereby rely only on more persistent features among multiple binary malware samples. However, more accurate classifiers will definitely benefit our system. 3: We would also like to experiment with unsupervised learning approaches in the training phase. Hereby, a machine learning algorithm might be able to select the ideal features that describe a given malware family best and weight the features correspondingly for the detection phase. 4: The malware detection might be improved by more sophisticated features that do not exploit periodicity alone but periodicity of communication sequences learned in the training phase, such as recurring three times a 20 minutes interval followed by a longer gap of 2 hours |  |  |

## 

## **Refid: 180, When Smartphones Become the Enemy: Unveiling Mobile Apps Anomalies Through Clustering Techniques**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| As future work, we shall dig deeper into feature selection approaches to improve the performance of the approach with the lowest intensity anomalies of type E2 | one needs to define in advance the number of K clusters to identify, which in principle is completely unknown, especially when no labeled data is used |  |

## 

## **Refid: 183, Radio Frequency Traffic Classification Over WLAN**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 185, Identifying Encrypted Malware Traffic with Contextual Flow Data**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| There are several research directions that could extend and improve on our work. Longer-term behavior could be observed, making it possible to use an FFT data feature to detect periodic communication patterns [38]. Training data could be extended to include honeypots and malware observed in the wild. Utilizing additional content-aware data features, such as SLADE [19], could be useful on non-TLS flows that are encrypted at the application layer. Like all vertical correlation systems, it could be extended by hybridizing it with a horizontal correlation system, such as one that analyzes the communication graph |  | threat detection on encrypted traffic |

## 

## **Refid: 188, Robust Network Traffic Identification with Unknown Applications**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In the future, we will develop practical software to facilitate unknown information extraction and apply it to real-time traffic classification. |  | threat detection on encrypted traffic |

## 

## **Refid: 195, Passive Classification of Wi-Fi Enabled Devices**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## **Refid: 196, A Modular Machine Learning System for Flow-Level Traffic Classification in Large Networks**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| We also plan to develop an automatic feature selection method that focus deliberately on improving stability. Our preliminary results show that by removing tos and tosnumbyte, we decrease the runtime complexity by 13.3% for TCP and 14.3% for UDP. In addition, the TCP flow/byte error rates are reduced by 16%/21% and the UDP error rates remain unchanged. We shall further refine this feature selection method | noisy data |  |

## **Refid: 199, Network planning tool based on network classification and load prediction**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Ongoing simulation and modeling work are carried out in order to prove that it is possible to enhance the resource allocation and the quality of service when we exploit these two algorithms. Due to lack of space, more simulations and evaluation results of our planning tool will be presented also on a forthcoming work. We also work on integrating this tool on an open source network simulator to help them to generate realistic optimal topologies taking benefits from our two algorithms. Finally, in the dataset analysis we will study correlation between different base station CDR sets. |  |  |

## **Refid: 202, Iterative-tuning support vector machine for network traffic classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| For future work, we propose to integrate the SVM classification approach based on flow-features with a heuristic approach based on the dynamic application characteristics to identify P2P traffic more accurately in the future work. and evaluation results of our planning tool will be presented also on a forthcoming work. We also work on integrating this tool on an open source network simulator to help them to generate realistic optimal topologies taking benefits from our two algorithms. Finally, in the dataset analysis we will study correlation between different base station CDR sets. | limitations faced by port-based schemes. |  |

## 

## **Refid: 208, CluClas: Hybrid clustering-classification approach for accurate and efficient network classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| A future direction of this research is evaluating our approach over different types of clustering and classification methods. Developing a theoretical proof also is left for future work. | limitations faced by port-based schemes. |  |

## 

## **Refid: 213, An Effective Network Traffic Classification Method with Unknown Flow Detection**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Since the flow label propagation is independent to classification algorithms, in the future, we can use it as a pre-processing step with any supervised methods in order to increase the size of the supervised training set. | limitations faced by port-based schemes. |  |

## **Refid: 217, TCFOM: A Robust Traffic Classification Framework Based on OC-SVM Combined with MC-SVM**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In future work, it is necessary to extract significant features from small flow or from the several packets when a flow began. In addition, the kinds of classified traffic are not many, but it is sufficient to verify that TCFOM can classify the traffics of P2P file sharing and web. |  | With statistical features in this paper, it is hard to classify the different applications which utilize the same techniques such as emule and BT. Thunder software is very popular in China and it supports the technologies of both emule and BT. It is required to extract the Thunder traffic features to strengthen TCFOM in future. The precision of identifying WOW traffic isn’t high, and we will extract WOW features to classify WOW traffic and Emule traffic. |

## 

## **Refid: 219, Network traffic classification using AdaBoost Dynamic**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| As future works, we wish to investigate what is the amount of test samples to keep the same level of accuracy. This analysis is important because it will be possible to build faster models with less data. Two questions are still open: 1) How to improve the generalization of the source IP? Only the destination IP was generalized, and all learned models suffer from the lack of transferability - the capability to learn from one network and transfer this learned model to a different network. 2) Does any geographic information help in algorithm accuracy? This feature was not tested in this work, because the GT data set was anonymized, but we wish to investigate if the use of any geographic data helps in the accuracy. |  |  |

## 

## **Refid: 222, A Biologically-Inspired Approach to Network Traffic Classification for Resource-Constrained Systems**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In the future, the training and classification time of the algorithm could be improved with the use of a k-d tree or Bloom filter data structure. Furthermore, the algorithm is inherently data parallel and can benefit from GPUs to speed up the training and classification times. |  |  |

## 

## **Refid: 232, Comparative study of a Hybrid Model for network traffic identification and its optimization using Firefly Algorithm**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Future work will follow similar directions, performing more tests on different data sets in order to continue to evaluate the robustness of the model. We are also interested in optimize the model not only for the correctness rate, but for multiple rates simultaneously, using a multi-objective optimization algorithm such as the Multi-objective Genetic Algorithm (MOGA) |  |  |

## 

## **Refid: 235, Network data classification using graph partition**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  | We faced the lack of standard web categorization method while carrying out our work. Though there exists few public web directories which maintained using crowd sourcing methods, the frequency of updating those directories are not enough to cover rapidly expanding World Wide Web. In order to use the full power of network partitioning there should be a frequently updating, standard directory service to label and categorize URLs |  |

## **Refid: 237, An approach for classification of network traffic on semi-supervised data using clustering techniques**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## **Refid: 240, Multi Level Statistical Classification of Network Traffic**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Our future work is to extend the proposed technique by enhancing the ground truth information and testing at realworld environment |  |  |

## 

## **Refid: 243, Network traffic classification in encrypted environment: A case study of Google Hangout**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Although the model showed to be very promising, reaching high levels on the performance metrics being investigated, it can still be improved, specially for the Mail class. To do so, new experiments could be conducted inspired by the divide-and-conquer strategy [32], where a complex task is divided into simpler (smaller) subtasks that together solve the main problem. It can be performed by dividing the data set into smaller clusters and then create a specific model for each cluster. Growing Hierarchical Self-Organizing Maps [33] could be used to divide the data set and Extreme Learning Machine to conquer each subset. The problem of imbalanced data set can also be better tackled with the Synthetic Minority Over-sampling Technique (SMOTE) [34], technique that shows to be very effective [30]. Besides that, a better investigation could also be performed on different data sets and to compare the results with other network traffic classification models. A deeper analysis of the model should be performed on high throughput networks (e.g., 10Gbps, 100Gbps) with the aid of Apache Hadoop [35], a framework that allows the distributed processing of large data sets across clusters of computers. |  |  |

## 

## **Refid: 248, Adaptive framework for network traffic classification using dimensionality reduction and clustering**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| More research is needed to find the most generally usable algorithms for each phase in the architecture. In addition, log data tends to be high in volume, so performance issues might become a problem. For dimensionality reduction the number of dimensions is not trivial. Also, the number of clusters must be determined depending on the chosen clustering algorithm. Real-time functioning requires changes in preprocessing and limits the dimensionality reduction options. For this purpose, PCA might be a good method, since projection of new points into lower dimensions is simply a matter of matrix multiplication. However, the limitations mentioned previously still apply |  |  |

## 

## **Refid: 249, Deep neural network based malware detection using two dimensional binary program features**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 251, Two-phased method for identifying SSH encrypted application flows**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Other plans for the future are to generate more extensive dataset and extend this analysis to other encrypted tunnels | The online performance of the classifier is dependent on the training dataset in other words how widely different behavior types of the applications have been captured and taught to the classifier. Therefore a more extensive training dataset would provide more reliable online classification.  Our method requires 1000 packets before the first classification of a tunneled application. With a packet size of 1.5 kbytes almost 1.5 Mbytes can be transferred without identifying the application. Especially in the security context it is important to identify threats quickly. This can be achieved by using smaller flow sample sizes |  |

## 

## **Refid: 253, Early detection of VoIP network flows based on sub-flow statistical characteristics of flows using machine learning**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Based on our experience of building a number of off-line classifiers we have developed a number of online network traffic classifiers based on Machine Learning and heuristics techniques. These classifiers will be deployed in our university network and the experience gained will be utilize to further refine our classifiers. Our group is in the process of obtaining more datasets from various sources which we intend to use to test our classifier for robustness. |  |  |

## 

## **Refid: 254, A Parallelized Network Traffic Classification Based on Hidden Markov Model**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Our ongoing work is (1) Improve our model’s accuracy of individual application; (2) Improve our model’s flexibility, enlarge the application range by adding typologies, such as SSH, to our model. (3) Comparing the performance with other classifiers |  |  |

## 

## **Refid: 260, Unknown malware detection using network traffic classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| For future work, we intend to extend the research toward transfer learning techniques to improve detection from untrained network environments, evaluate the proposed methods and models on mobile network traffic, test the proposed methods for malware family clustering, and finally adjust the method for online detection for high bandwidth networks. |  |  |

## 

## **Refid: 261, GPU-oriented stream data mining traffic classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| As future work, we intend to carry out an extensive analysis of the tree build time, and the optimization of the build process. Additionally, we aim to evaluate the computational performance of the traditional classification techniques into GPU and analyze other statistical metrics. Finally, we plan to evaluate the impact of varying the number of blocks and threads in the GSDT performance. |  |  |

## 

## **Refid: 264, Classification in dynamic streaming networks**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| For future work, we will investigate the pros and cons of our incremental methods by conducting comparisons with state-ofthe-art algorithms on more real-world dynamic networks, and explore the theoretical relationship between the user-defined variables (i.e., window size, edge extraction threshold) and the classification performance of the proposed algorithms. We would also like to investigate how to extend the proposed methods to study the problem of supervised learning on dynamic graphs with insertions, deletions and modifications of nodes/edges. |  |  |

## 

## **Refid: 267, Implementation of network traffic classifier using semi supervised machine learning approach**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| If we utilize the proper feature selection technique than it may be possible that less labeled instances will be required to achieve higher accuracy with less computational time. Overall, this approach achieved better results. In the future, this approach could become an excellent tool to classify network traffic |  |  |

## 

## **Refid: 269, Classification of network traffic in LAN**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| It is a good idea to stay updated to this kind of attack and study those types of attack due to which the real network is facing problems day by day and can reduce many network problems like Bandwidth utilization etc. Optimizing the parameters present in the algorithm reduces the training time. More reduction techniques may be referred to get valuable features in future and developed a new detection model with higher performance. |  |  |

## 

## **Refid: 282, BalancedBoost: A hybrid approach for real-time network traffic classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 286, Network Traffic Classification Using Tri-training Based on Statistical Flow Characteristics**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## **Refid: 292, A novel semi-supervised approach for network traffic clustering**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In the future, we plan on investigating a fully unsupervised data sampling approach for traffic clustering  Another important issue is the number of clusters. In the manner of unsupervised learning, K is unknown in advance. The common approach is to set it to a large number and to have multiple clusters mapping to each application class. For the cluster to class mapping, labeled data can be used [12]. We plan to explore the effectiveness of utilizing side information for this task in the future |  |  |

## 

## **Refid: 294, Incremental high throughput network traffic classifier**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In future, the architecture will be implemented in an upgraded NetFPGA version to improved the classification performance. |  |  |

## 

## **Refid: 297, Investigation of machine learning based network traffic classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Our future work will focus on applying the classification system described in this paper to real SDN traffic data which could have very different characteristics than the dataset we used. More importantly, how to leverage traffic and application awareness enabled by machine learning techniques in wireless SDN networks to enhance traffic engineering, SDN forwarding rules, network slicing, and flow QoS management is a promising research topic. |  |  |

## **Refid: 304, EnClass: Ensemble-Based Classification Model for Network Anomaly Detection in Massive Datasets**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 306, Network Traffic Classification Using Correlation Information**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In the future, we will work on developing new methods for flow correlation analysis.  Considering the applications using UDP are growing, we plan to further evaluate the proposed approach using UDP flows in the future. |  |  |

## 

## **Refid: 308, A self-adaptive network traffic classification system with unknown flow detection**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  | Given the initial centers and information gain weight of each feature, the semi-supervised k-means algorithm aims to partition the traffic flows into k clusters. The setting of the parameter is always a significant challenge for a traffic classification method that applies machine learning. |

## 

## **Refid: 311, Dynamic online traffic identification scheme based on data stream clustering algorithm**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Our future work is to get the ideas of How to generate dataset which can simulate the dynamic feature of network traffic and how to find an efficient way to satisfy classification requests of both temporary and historical. |  |  |

## 

## **Refid: 313, Identification in Encrypted Wireless Networks Using Supervised Learning**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Detecting abrupt change is future work |  |  |

## 

## **Refid: 317, Comprehensive Analysis of Network Traffic Data**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In the future, we aim to keep improving the classification accuracy by optimizing the hyper-parameters for classifiers and introducing advanced techniques of combining classifiers. Moreover, we intend to introduce novel performance metrics for evaluation. We would also like to improve the visual presentation of the data set despite of its complexity and large volume |  |  |

## 

## **Refid: 323, Traffic Identification in Semi-known Network Environment**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 328, Metrie learning with statistical features for network traffic classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 333, Robust Network Traffic Classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| We plan to extend this work in the future and address the problem of changing class distribution by developing new strategies for system updates and classifier retraining. One idea is to count the flows of any known classes recognized by semi-automatic identification during a system update. If the number increases, this indicates class distributions of the corresponding known classes have changed and the traffic classifier should be retrained. In other words, when changed ZHANG et al.: ROBUST NETWORK TRAFFIC CLASSIFICATION 1269 class distributions or new classes are detected, the system update will be triggered |  |  |

## 

## **Refid: 336, P2P traffic classification using clustering technology**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## **Refid: 342, SMILER: Towards Practical Online Traffic Classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| For example, with very short flows (less than three packets) or lots of disordered packets, SMILER’s accuracy might be reduced, which implies that approaches that are more robust should be designed in future. Our future work also includes improving the classification performance of SMILER and evaluating on various network environments |  |  |

## 

## **Refid: 353, WekaTIE, a traffic classification plugin integrating TIE and Weka**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| our plans for immediate future work are to ease the whole procedure, seeking for mechanisms to automatize the model generation, and avoiding, as possible, the need to use Weka GUI as an intermediate step. Moreover, we plan to further test the possibility of using wekaTIE plugin as an online traffic classification tool. |  |  |

## 

## **Refid: 354, Harvesting unique characteristics in packet sequences for effective application classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| However, the proposed system is able to discover the effective PPS signature for all types of Skype media traffic. Furthermore, we present PPS signatures for another two applications as well in TABLE.II. Due to the limited space, more experimental evaluation for the APSC system will be presented in our future work. |  |  |

## **Refid: 357, A Heuristic-Based Co-clustering Algorithm for the Internet Traffic Classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Based on this observation, in the immediate future we will investigate the following three questions in more details: Do we need to filter our some unknown/known flows before classification? This is because, for example, we have some known flows that are identified as attacks and for those flows with attacks labeled, we will exclude them when doing classification. Is there any similarity among applications on individual days; in particular the daily traffic and monthly traffic? We have investigated the hourly traffic on two individual days and we will continue to compare the similarities of daily traffic and monthly traffic in that we can construct some new statistical models for traffic classifications. Is there any possible reverse flows in the identification framework? We have found that for some flows to be labeled as unknown is simply because the direction of the flows is wrong when creating flows based on individual packets. In such a case the applications can be easily identified as long as we reverse the direction of the flows. |  |  |

## **Refid: 360, Network Traffic Classifier With Convolutional and Recurrent Neural Networks for Internet of Things**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Being the deep learning architectures such a fruitful source of new models, we consider, as future work, to experiment with new applications and variants of the CNN and LSTM models |  | In order to train the method, we have used more than 250,000 network flows which contained more than 100 distinct services. As an additional challenge, the frequency distribution of these services was highly unbalanced |

## 

## **Refid: 363, Network classification using adjacency matrix embeddings and deep learning**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 375, Network traffic classification using machine learning algorithms**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 392, Internet traffic classification based on flows' statistical properties with machine learning**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| The terminology is one of the issues the future researchers have to deal with. Moreover, an important problem is the lack of shareable test data input along with re-labeled flow objects to be used as a reference. Privacy and ownership constrains when disclosing the data make the mission of researchers hard in their attempt to improve the way traffic is classified. They actually encourage, inside dedicated communities, the development of tools that annotate the training data with a class. They also find important the capacity of building scalable, parallel traffic classifiers that will be able to deal with the huge amount of data that are coming in the context of networks being more and more powerful. |  |  |

## 

## **Refid: 400, A comparison of supervised machine learning algorithms for classification of communications network traffic**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In current research we are using these results as the basis for designing dynamic network classification techniques that can respond in real time to changes in the traffic profile. | it required a significant processing power |  |

## 

## **Refid: 412, An entropy based encrypted traffic classifier**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| As a future work, we have plan to investigate our proposed model on a wider range of applications and heterogeneous datasets from various networks. We also aim at analyzing protocol’s individual signature from entropy and cross-validate its consistency with clustering algorithms. Finally, we plan to apply this approach to reveal signatures for network intrusion detection system. That would make it easy to infer if the encrypted traffic is benign or something that should be investigated further |  |  |

## 

## **Refid: 418, Improved classification of known and unknown network traffic flows using semi-supervised machine learning**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  | However, these approaches can only predict predefined classes found in the training data. Unsupervised learning approaches [7, 8] classify from clusters of unlabelled training flows. While using unlabelled data means they can handle known and unknown classes, mapping clusters to classes remains a key challenge |

## **Refid: 419, Hybrid multi-objective optimization approach for neural network classification using local search**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 433, A preliminary investigation on the identification of peer to peer network applications**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Future work should investigate optimization of P2P application identification results by using another algorithms such as SOM, Fuzzy C-Means, improving current algorithms or developing a new algorithm. In addition, the FPR was high for the training and testing for UTorrentDS and BitTorrentDS data sets for most the DM algorithms. Generating a bigger dataset and implementing these algorithms over this set may help us overcome these drawbacks. Finally, other traffic analysis approaches can be applied to see their effectiveness on this task |  |  |

## 

## **Refid: 439, Real-time multi-application network traffic identification based on machine learning**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Future work is aimed to improve constantly the rationality and validity of ”time window” feature extraction for the unique characteristics of network application, and to make the network flow identification methods based on SVM more stable and more robust. |  |  |

## 

## **Refid: 452, Analysis of Early Traffic Processing and Comparison of Machine Learning Algorithms for Real Time Internet Traffic**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| It is also suggested to see the effect of flooding traffic by different tools like WAR-FLOOD, Trinoo, TFN and metasploit scripts will make this real time environment more interesting using ML classifiers. Also other optimum and orthogonal statistical flow discriminators can be tried. |  |  |

## **Refid: 453, On addressing the imbalance problem: A correlated KNN approach for network traffic classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  | A significant challenge to the classification performance comes from imbalanced distribution of data in traffic classification system |

## **Refid: 489, TrafficS: A behavior-based network traffic classification benchmark system with traffic sampling functionality**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## **Refid: 547, Better network traffic identification through the independent combination of techniques**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| A future work is the fine-grained evaluation of the segmentation sizes between 100 and 1000 bytes along the different network scenarios. Another future work is the proposal of a combination methodology that supersedes the presented combinations by providing results closer to the perfect combination (theoretical best performance) in the backbone scenarios. |  |  |

## 

## **Refid: 548, An accurate traffic classification model based on support vector machines**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Further research and improvement on SVM multi‐ classifier are ongoing. |  |  |

## **Refid: 549, Network Traffic Classification using Genetic Algorithms based on Support Vector Machine**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 560, A novel semi-supervised learning method for Internet application identification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## **Refid: 569, Ensemble network traffic classification: Algorithm comparison and novel ensemble scheme proposal**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 577, Waterfall: Rapid Identification of IP Flows Using Cascade Classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| The future research could focus on a more detailed comparison between classifier fusion and classifier selection in context of traffic classification. Another interesting problems are the optimal selection of training instances for the classification modules (in accordance with their criteria), and the proper sequence of modules in the cascade. |  |  |

## **Refid: 583, Network traffic classification based on ensemble learning and co-training**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Although this classification model performs much better than traditional methods, many problems are left for further work, such as 1) in the ensemble learning mode, how to maintain the diversity of classifiers to the great extent, 2) and in the co-training mode, how to utilize the unlabeled data more effectively. We believe it is worthy of being studied in the future. |  |  |

## 

## **Refid: 598, High-Performance Internet Traffic Classification Using a Markov Model and Kullback-Leibler Divergence**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 603, Network traffic classification via non-convex multi-task feature learning**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In the future work, we will extend the work to a general parallel computing framework due to the inherent parallelization form of (9), the d independent optimization problem could be solved simultaneously. The considerable platform is GPU or distributed |  |  |

## 

## **Refid: 610, SmoteAdaNL: a learning method for network traffic classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| SmoteAdaNL can be further improved by considering more specific situations. With respect to data sampling, the method of handling class overlapping region can be further improved from the data level, such as using data cleaning method (Wang et al. 2014). Additionally, this paper only works on wired network traffic without considering the wireless mobile networks (Ikeda et al. 2013). With the imperative evolution from IPv4 to IPv6, the applications in mobile IPv6 networks (Tao et al. 2012, 2014) experience a rapid growth, we will facilitate the network traffic classification problem on the mobile traffic traces collected from IPv6 network as our future work |  |  |

## **Refid: 611, A class-oriented feature selection approach for multi-class imbalanced network traffic datasets based on local and global metrics fusion**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In pattern recognition field, there are some related studies, such as feature similarity index [39], discrimination capability of coding vectors [40] and information projecting principle based learning algorithm [41]. These are mainly designed for image processing or image classification, and cannot be directly used for network traffic classification task. However, some ideas in these references are worth to be researched. We will study how to apply these related works from other fields into network traffic classification in future |  | Multi-class imbalance problem  Data Drift Problem |

## 

## **Refid: 618, Encrypted Traffic Classification Using Statistical Features**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  | The first major challenge ahead of this method is to analyze the payload of every single network packets, which requires heavy processing efforts. In addition, contents of encrypted packets are ambiguous for these methods. These restrictions make them unusable for such cases. Respecting the privacy of the users is another imp |

## 

## **Refid: 644, Effective Packet Number for 5G IM WeChat Application at Early Stage Traffic Classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| There is still gap for further research in the early Internet traffic classification. A new method should be developed to select effective packet numbers for 5G WeChat application early stage traffic identification, while selecting more packets for Internet traffic classification increases computational complexity while minimum features will decrease classification accuracy of machine learning classifier so that more models should be developed that show how many packets should be used for accurate IM application traffic classification. |  |  |

## 

## **Refid: 660, Machine learning algorithms for accurate flow-based network traffic classification: Evaluation and comparison**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Our future work includes investigating the classification performance of unsupervised ML algorithms following a similar methodology, and using a large amount of reliable data. The classification of P2P traffic is a problem of high relevance. We will investigate the effects of including different individual P2P applications in the training and testing data sets on the accuracy of the classifier |  |  |

## **Refid: 663, Using clustering to improve the KNN-based classifiers for online anomaly network traffic identification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Applying the MLBG clustering algorithm to the sampling dataset in advance can further improve performance and greatly reduce the training time and online classification time. |  |  |

**Refid: 664, NTCA: A High-Performance Network Traffic Classification Architecture**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| We are pursuing this work in three points. The first is focusing on applying the hybrid method to measure P2P protocols. The second is searching more effective flow-level features. The third is applying our traffic classification architecture to the actual system for the real-time traffic. |  |  |

## 

## **Refid: 676, Internet traffic classification based on flows' statistical properties with machine learning**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  | The terminology is one of the issues the future researchers have to deal with. Moreover, an important problem is the lack of shareable test data input along with re-labeled flow objects to be used as a reference. Privacy and ownership constrains when disclosing the data make the mission of researchers hard in their attempt to improve the way traffic is classified |  |

## 

## **Refid: 689, On Evaluating Multi-class Network Traffic Classifiers Based on AUC**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In the future, we can perform the following work: 1. Use more classifiers to verify our method. 2. Find more proper metrics for evaluating multi-class network traffic classifier. Based on those, we can select the most suitable method for network traffic classification |  |  |

## **Refid: 691, An SVM-based machine learning method for accurate internet traffic classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In our future work, we intend to combine the supervised and un-supervised machine learning methods, as well as using feature parameters obtainable early in the traffic flow for fast and accurate Internet traffic classifications. | One of the disadvantages of SVM-based and other supervised machine learning method is the requirement on a large number of labeled training samples. Moreover, identifying the traffic after the network flow is collected could be too late should security and QoS intervention become necessary in the early stage of the traffic flow |  |

## 

## **Refid: 704, Metric Learning With Statistical Features For Network Traffic Classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 707, Network Traffic Classification Model Based on MDL Criterion**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## **Refid: 710, Internet Traffic Classification Demystified: On the Sources of the Discriminative Power**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  | : (i) key feature selection, (ii) finding the best algorithm(s) for traffic classification, and (iii) obtaining representative data sets for training and testing machine learning algorithms |

## **Refid: 718, Early Traffic Classification Using Support Vector Machines**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Maximum captured packet size was limited to 200 bytes, so some patterns may have gone undetected, which led to an inaccurate payload based classification. This was a limitation of available captured data. We would like to get full packet 0 0.2 0.4 0.6 0.8 1 msn ssdp ares smtp smb nbns pop3 ssl gnutella dns http bittorrent Overall Classification accuracy Traffic type SVM classification accuracy by traffic type 5 packets per flow direction, 45 flows to train C=512 γ =2 Centroid SVM Figure 8: Centroid and SVM accuracy per traffic type size traffic traces to improve the L7-filter pre-classification phase. We have not filtered flows initiated before the start of the captured data. Most probably these flows were not detected by payload analysis so they were probably excluded of clustering analysis. Detecting them and filtering out, may lead to better results. We have not filtered initial SYN packets for TCP connections either, so for TCP flows the first packets do not provide relevant information since they have zero size. This obviously does not apply to UDP traffic. We are not taking care of unordered packets. Captured data may have a small amount of unordered packets that may lead to inaccurate payload classification. We do not analyze the case of unknown traffic. In this study every pre-classified flow is assigned to one of the known traffic classes. Other sources of captured data traffic should be included in future work. Although it is known that the SVM technique is fast, the computational cost of both tested methods was not analyzed in this work. Besides these limitations and restrictions the methodology proved powerful, and results are promising. |  |  |

## **Refid: 724, Challenging Statistical Classification for Operational Usage: The ADSL Case**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| As future work, we would like to devise a strategy to select features that would be immune when used in cross-site studies. A possible solution could be to use per application features instead of a shared set for all applications. |  |  |

## **Refid: 725, MINETRAC: Mining Flows for Unsupervised Analysis & Semi-supervised Classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## **Refid: 727, Rapid Identification of Skype Traffic Flows**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Our future work will be to identify additional features, particularly based on interarrival time, to identify Skype traffic using shorter windows than 5 seconds. We also plan to investigate whether these particular features generalize to other codecs used in the Skype application. |  |  |

## 

## **Refid: 729, High Throughput and Programmable Online Trafficclassifier on FPGA**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In the future, we will study the tradeoffs between performance and cost of the two architectures. We will also explore the potential of our work in a broader scope of high speed classification problems. |  |  |

## 

## **Refid: 738, Timely and Continuous Machine-learning-based Classification for Interactive IP Traffic**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Our work can be extended in a number of future research directions: • expanding our approach toward the recognition of different or new applications (e.g., video streaming); • identifying how different sub-flow sizes affect classification accuracy, classification timeliness, classification speed, and the stability of results for continuous classification, as well as characterizing the optimal sub-flow size for applications; • evaluating the impact of different traffic mixes and different ML classification and clustering algorithms on classification accuracy; • evaluating the scalability of our approach to classify a large number (for example, hundreds) of applications simultaneously; • evaluating the portability of classification models (when being applied in different network environments); • evaluating the stability of classification accuracy and characterizing the sensitivity of different ML algorithms when using packet sampling, employing different features, and in the presence of network perturbations, such as packet loss, delay, and reordering |  |  |

## 

## **Refid: 741, On Botnet Behaviour Analysis Using GP and C4.5**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Future work will explore what other flow features can be employed in botnet behavior analysis and their effects in terms of lowering the false alarm rates. |  |  |

## 

## **Refid: 746, KISS: Stochastic Packet Inspection Classifier for UDP Traffic**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| We assume packets belonging to the same flow/endpoint are exposed to the KISS engine, so that after digesting packets, a classification decision is taken and a new observation window begins. Therefore, several classification decisions are possibly taken for a single flow or endpoint. In this paper, we consider independent classifications, so the same flow/endpoint can be classified differently at each window. Notice that some reconciliation algorithm can be easily designed to increase the accuracy of the classification by considering the set of classifications involving the same flow or endpoint, e.g., adopting a majority criterion. We leave this issue to future work |  |  |

## 

## **Refid: 756, Detecting Advanced Persistent Threats Using Fractal Dimension Based Machine Learning Classification**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Future work includes testing of the data series using other machine learning algorithms like Naive Bayes, Support Vector machines, Decision Trees and Boosted classifiers to validate the performance of the proposed methodology. Moreover, using more feature vector, other fractal dimensions like Self-Similarity dimension, Hausdorff dimension, Information dimension and variance fractal dimension can be explored for classification of advanced malicious data |  |  |

## 

## **Refid: 759, Statistical Traffic Classification by Boosting Support Vector Machines**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| For future work we would like to analyze the computational cost of SVM classification, and the penalty added by the inclusion of the boosting technique. Also it will be interesting to analyze the classification accuracy considering only one way of the flow, something very common on a large ISP with multiples network connections. The unknown traffic analysis would be also an interest point to address in future works. |  |  |

## **Refid: 762, Traffic Classification Using Visual Motifs: An Empirical Evaluation**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| We leave as future work an examination on how the techniques can be used to distinguish between different protocol behaviors of the same application (e.g, the file transfer vs. interactive mode of SSH, or streaming video and audio over HTTP). | A limitation of this work is its susceptibility to evasive maneuvers [1, 16], where a user attempts to disguise her actions by morphing or padding her packets to look like a different protocol. Techniques for mitigating such attacks remain as future work. A more practical limitation is the fact that we assume that sessions can be separated from each other. Clearly, while this assumption holds for cleartext traffic or even for separating traffic employing application layer encryption (say using SSL), the assumption does not hold in cases where traffic from disjoint protocols is multiplexed over a single encrypted tunnel—as is typically done when using VPNs. Lastly, the approach we take is only applicable to TCP traffic. That said, since TCP dominates UDP traffic (e.g., 86% of the dartmouth data was TCP traffic), we argue this limitation may not be that significant in practice. |  |

## 

## **Refid: 763, Classifying SSH Encrypted Traffic with Minimum Packet Header Features Using Genetic Programming**

## 

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Future work will follow similar lines to perform more tests on different and/or larger data sets in order to continue to test the robustness of the classifiers as well as exploring the appropriateness of other machine learning algorithms. We are also interested in defining a framework for generating ‘good’ training data sets, where this might include combining training data from multiple independent sources. Evaluation under other encrypted applications as well as exploring the possibilities for integrating our approach with approaches employing host based behavior are also of interest. |  |  |

## 

## **Refid: 778, A Functional Approach to Scanner Detection**

## 

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 782, Entropy Based Adaptive Flow Aggregation**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## **Refid: 785, Online Internet traffic monitoring system using spark streaming**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In future, we will implement collectors to capture packets from switches through port mirroring so that our system can analyze all the traffics passing through monitored networks. Finally, we will test its performance in practice and compare it with some traditional single server systems in terms of scalability and reliability. |  |  |

## 

## **Refid: 791, Online Traffic Classification Based on Co-training Method**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 799, Comparative analysis of five machine learning algorithms for IP traffic classification**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In this research work, internet traffic dataset has been developed by considering packet flow duration of 2 minutes for each application which is still very large. This flow duration can be further reduced to make analysis more real time compatible. Secondly, internet traffic can also be captured from various different real time environments such as university or college campus, offices, home environments etc. This internet traffic dataset can be extended for many other internet applications |  |  |

## 

## **Refid: 809, Cascaded classifier for improving traffic classification accuracy**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 810, Performance Comparison of Four Rule Sets: An Example for Encrypted Traffic Classification**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| For future work, we are interested in investigating our approach for other encrypted applications such as Skype traffic and test its robustness with more data sets |  |  |

## 

## **Refid: 816, A Statistical-Feature ML Approach to IP Traffic Classification Based on CUDA**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| For future work, we are interested in investigating our approach for other encrypted applications such as Skype traffic and test its robustness with more data sets |  |  |

## 

## **Refid: 818, A statistical-feature-based approach to internet traffic classification using Machine Learning**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In our future work, complete tests about the current approach are badly needed (e.g., using adequate data samples for training for the sake of the principle of fairness in KNN algorithm). Furthermore, we’ll apply numbers of refinements to improve the classification results such as meticulous process of feature selection and optimization of the classifier model. This paper is just one of those that have applied ML techniques in traffic classification. Further investigations like using unsupervised ML techniques to classify unknown applications (not marked when training phase at all) and classification techniques on the basis of multi-flow topology [19] are also on the schedule if needed. |  |  |

## 

## **Refid: 829, A modular two-layer system for accurate and fast traffic classification**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## **Refid: 835, Internet traffic classification using MOEA and online refinement in voting on ensemble methods**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 837, Compact Hash Tables for High-Performance Traffic Classification on Multi-core Processors**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In the future, we plan to explore hashing for the merge stage of our approach as well. We also plan to do a thorough comparison among various types of platforms, including multi/many-core processors, GPUs, and FPGAs. |  |  |

## 

## **Refid: 841, SeLeCT: Self-Learning Classifier for Internet Traffic**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| The choice of which features to consider is a matter of optimization and several works in the literature have proposed and investigated possible alternatives. Our choice stems from the following intuitions: (i) keep the feature set limited, (ii) include generic layer-4 features that can be easily computed, and (iii) use features that can be collected during the beginning of a flow so that we can classify flows in real-time (i.e., minimize the time required for identification). It is out of the scope of this paper to compare and choose which are the most suitable features to use |  |  |

## 

## **Refid: 849, Sub-flow packet sampling for scalable ML classification of interactive traffic**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In future work we aim to dynamically adjust these parameters to achieve a desirable sampling rate and optimise the memory and CPU usage. |  |  |

## 

## **Refid: 852, Evaluating shallow and deep networks for secure shell (ssh)traffic analysis**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In our future work, we will focus on data gathering, labeling and asses that labeled feature vectors using machine learning to find out the effectiveness of them towards robust traffic classification. Employing deep learning approaches on raw network traffic traces by using [27] approach will be remained as another future direction of our work. |  |  |

## 

## **Refid: 854, How many packets are most effective for early stage traffic identification: An experimental study**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 855, BotDetector: A robust and scalable approach toward detecting malware-infected devices**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  | BotDetector uses information of the HTTP header fields; hence, the system cannot react when traffic is encrypted like HTTPS Also, HTTP protocol is the primary protocol that malware use, but the system also cannot respond to malware using other protocols. For example, it was difficult to apply to malware that often uses UDP protocols. Another limitation of BotDetector is that malware developers can change the HTTP headers to evade detection; i.e., the traffic originated from malware can mimic the traffic originated from a browser. Although the case is not major today, such evasion could become common in future, at which point, we need to change the feature extraction and classification model. However, we believe that the fundamental idea behind this work – finding useful features automatically – remains beneficial to discover invariants that could be used to detect malicious activities |  |

## 

## **Refid: 859, Online Internet Traffic Measurement and Monitoring Using Spark Streaming**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| As an ongoing project, we are implementing the proposed system in an experiment environment that the collector can only capture inbound and outbound packets of a single computer. In the future, we will capture packets from switches through port mirroring, so that our system can analyze all the traffics in monitored networks. Finally, we will test its performance in practice, and compare it with some traditional single server systems from both scalability and reliability. |  |  |

## 

## **Refid: 866, Multi-layer Anomaly Detection for Internet Traffic Based on Data Mining**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In the future, detection can be more specific in analyzing the behaviors of the internet traffic instead of only identifying the basic fields, making the result more detailed and increasing its accuracy. |  |  |

## 

## **Refid: 870, A novel aggregated statistical feature based accurate classification for internet traffic**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| As a future work, the proposed method can be applied to the field of Big Data using Hadoop technology. The method can be extended by classifying real time packets efficiently. Focusing more on R programming concepts which is a programming language and software environment for statistical computing widely used by the statisticians and data miners for developing statistical software will helps new improvement in the field of traffic classification. |  |  |

## 

## **Refid: 873, Efficient Methods for Early Protocol Identification**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| A thorough performance analysis was made on both laboratory and real world traffic traces to examine the accuracy of our algorithms. The laboratory tests were carried out on actively collected data: in this case the ground truth is known without doubt, but the traffic patterns are less realistic than in real traces. However, we believe that the fact that our algorithms use only the first 16 bytes of the flows makes our method less sensitive to such errors (apart from the fact that the TP and FP rates are highly dependent on the composition of the whole data set). While during laboratory experiments only a few P2P protocols were taken into account, we also showed how the proposed techniques scale up to handle a large number of various protocol types that may appear in real world traffic. To this end, the proposed methods were also validated on a much more diverse data set collected in the network of a large European ISP, and achieved surprisingly good accuracy for the majority of the protocol types on these real traces. We also considered the problem of handling previously unseen traffic types, and showed training methods to deal with such problems. | In this section we discuss the limitations of our earlyclassification approach, which need to be addressed via different techniques. First of all, since our methods purely depend on the payload of the network traffic, encrypted network flows cannot be handled at all. On the other hand, the vast majority of today’s Internet traffic is still unencrypted. Although the proportion of encrypted traffic will likely increase in the near future, it is expected that a significant portion of the total traffic will remain unencrypted, due to several reasons, for example, since encryption requires a lot of computational resources (hence quickly reduces the battery life of mobile devices) and is not necessary for most of the applications. Another issue we have to deal with is related to encapsulation when the content of a given application is embedded into another trusted protocol like HTTP. Several applications use this technique to go through firewalls. If a protocol is embedded into simple HTTP, despite the unencrypted data the proposed techniques would identify the carrying HTTP instead of the real applications. Our methodology in its current form is not prepared for such situations when the payload does not start with the protocol specific content. However, this can be solved by the segmentation of HTTP/HTML messages. To this end, the embedded content has to be searched for, extracted, and than the proposed technique can easily be applied for the separate segments. Furthermore, by their nature, our methods are vulnerable to countermeasures when a malicious user injects some arbitrary bytes into the beginning of the payloads to mislead the classifier. Fortunately, it is not the general case, and most of the users use the network protocols as they are. If, however, a malicious user modifies the payload, the accuracy of our methods may substantially degrade. This problem may be alleviated by looking for the most protocol specific parts of the payload during the training instead of using the first bytes of the flows. As we mentioned before, this issue is already present in case of UDPbased LimeWire traffic, where the first 16 bytes do not follow a well defined protocol specific pattern. This phenomena may have various reasons: the first bytes carry increasing sequence numbers, random identifiers, filenames, etc. However, if the training set is large enough and contains accurate class labels, only a slight effect on the accuracy of our approaches is expected to have. Finally, we also have to mention that in commercial products it is not uncommon to have hundreds of application classes to be handled. Although the most widely used protocols are considered in the experiments of this paper, the protocol portfolio will certainly change in the future, and new protocols should also be considered. We expect that our methods would scale up well to such problems and achieve similarly good performance in general. Nevertheless, with the number of traffic classes increasing, our methods might become more sensitive to the above mentioned limitations (e.g., if more and more new protocols start with random strings). In spite of the revealed limitations, the proposed techniques show sufficiently good performance even in real world environments. Furthermore, since they have extremely low computational and storage complexities and can easily be implemented even in hardware, they could serve as lightweight pre-classifiers that can efficiently label or even filter out wanted or unwanted traffic in the very early stage of a byte flow. If the methods are applied for prescreening, one can also consider a refined version of the algorithms where the classifiers also report the uncertainty of their label predictions (e.g., based on similar measurements as reported in Fig. 10), and further, more resource-intensive methods can be applied to deal with the hard cases where the uncertainty of our algorithms is high. |  |

## 

## **Refid: 882, Genetic optimization and hierarchical clustering applied to encrypted traffic identification**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| With regards to the future work, given that gains in performance were observed with the implemented two level hierarchical approach, it would be interesting to observe the effects of implementing additional layers of clusters. This, combined with larger training data sets could potentially lead to better results. Also, in this work SSH was chosen as an example of an encrypted traffic application. This same approach could also be used with other types of encrypted traffic, such as SSL, or Skype. Moreover, it is also of interest to identify applications within an encrypted tunnel, or to identify malicious behaviors hidden in a tunnel. |  |  |

## 

## **Refid: 890, Anonymity Services Tor, I2P, JonDonym: Classifying in the Dark (Web)**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| As future work we will investigate (i) hierarchical classification, (ii) comparison with other public labeled datasets (possibly also in an open-world assumption), should they become available, (iii) development of classifier fusion techniques for anonymous TC, and (iv) implementation of features and classifiers in the open-source TC platform TIE [47] to allow researchers to evaluate them on live traffic traces. |  |  |

## 

## **Refid: 892, Identifying Key Features for P2P Traffic Classification**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Our future work will be addressing first of all the extension of the set of features, taking into account different levels of aggregation (e.g. host-level and port-level) along with more complex operations on the counters, as these have been shown to enhance classification accuracy [12], [30]. We also want to improve the estimation of mutual information, for instance by employing non parametric methodologies [29], possibly in combination with specific feature selection schemes based on this estimation, like e.g. [20], which will enable a better assessment of the classification performance. |  |  |

## 

## **Refid: 895, Real-Time Classification of Multimedia Traffic Using FPGA**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Our future work includes applying the proposed solution to classifying more application types, integrating the FPGA design with the network interface on the development board and testing its performance under real-life network traffic |  |  |

## 

## **Refid: 903, Early classification of residential networks traffic using C5.0 machine learning algorithm**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| We are currently working on the integration of a flow monitoring exporter on a real home gateway device to conduct an online performance evaluation of our proposed MLA classifier. This task is technically challenging considering some hardware architectural issues (i.e. hardware accelerators that prevent packets observation at Linux kernel stage). We are also planning to extend our data collection process to more locations and network configurations. |  |  |

## 

## **Refid: 904, Hybrid Traffic Classification Approach Based on Decision Tree**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  | One of the biggest limitations for our hybrid approach is it fails to find the new (or unknown) applications. Unsupervised learning techniques have been studied recently to discover the new applications on the Internet, it, however, suffers from a large number of false positives. In order to address this issue, in the immediate future we will extend our approach to discover automatically new applications. The basic idea behind this is that we will apply the two decision tree classifiers based on source flow and destination flow in parallel, if both classifiers agree that an unknown flow is, for example, Gnutella, then the final classification output will be, to say, Gnutella. Otherwise if the two classifiers obtain contrary identification results, to say, one labels the unknown flow into Httpweb, while the other says it is BitTorrent. In this case we will leave the flow as unknown. This idea for new application discovery is reasonable and can be supported by our experimental results. During the evaluation, we found that the classification capability for the two individual classifiers is very similar. The same flow identified into two totally different applications by the two classifiers with similar classification capability might stand for a new application. Moreover, in the near future we will release and share a large size dataset collected on a free assess WiFi network to the public for comparing the performance of different classifiers in the academia, addressing one of the open issues mentioned in [6]. |  |

## 

## **Refid: 905, The Impact of Evasion on the Generalization of Machine Learning Algorithms to Classify VoIP Traffic**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In the future, we plan to test the proposed approach on different types of evasion attacks as well as on larger and more diverse network traces. | . |  |

## 

## **Refid: 919, Internet Traffic Classification Using Constrained Clustering**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  | . |  |

## 

## **Refid: 926, Hierarchical RBF Neural Network Using for Early Stage Internet Traffic Identification**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  | . |  |

## 

## **Refid: 928, Noise-resistant Statistical Traffic Classification**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  | . | This paper presented a real-world challenge that network traffic classification struggles to perform well in the presence of mislabelled samples in the training data. That is, when mislabelled traffic samples are present, conventional traffic classification methods cannot sustain their performance. We proposed a novel traffic classification method, noise-resistant statistical traffic classification (NSTC), which can identify noisy examples and tolerate suspected noisy samples |

## 

## **Refid: 930, Internet traffic classification based on associative classifiers**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  | . |  |

## 

## **Refid: 934, Practical machine learning based multimedia traffic classification for distributed QoS management**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Future work remains to evaluate the performance of DIFFUSE on (low-end) routers/gateways, analyse load effects of frequent adding/removal of flows and more comprehensively analyse the system’s classification accuracy, timeliness and robustness. We also plan to explore whether automatic (re)training of classifiers may be practically achieved using live IP traffic, and the degree to which noise (packet loss and jitter) in the live traffic negatively impacts on the system’s ability to recognise the same class of traffic in the future. | . |  |

## 

## **Refid: 938, Traffic classification using cost based decision tree**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  | . |  |

## 

## **Refid: 951, A hybrid heuristics-statistical peer-to-peer traffic classifier**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| For future work, our traffic classification system can be integrated with any network control system. Any proposed traffic control system should consider the nature of P2P applications which negatively affect the existing network architecture in term of the consumption of the upload bandwidth. Furthermore, the evaluation of proposed system was performed on offline traces. Future work should enhance this system to work with live traffic in online network to deeply assess the performance of online phase and study the effect of periodical retraining. The time between retraining and number of instances that should be used need further analysis for J48 algorithm or any decision tree algorithm. | . |  |

## 

## **Refid: 958, Experiments on detection of Denial of Service attacks using Naive Bayesian classifier**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  | . |  |

## 

## **Refid: 959, Seq2Img: A sequence-to-image based approach towards IP traffic classification using convolutional neural networks**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Future work will be focused on introducing more raw feature sequences to improve the accuracy. | . |  |

## 

## 

## **Refid: 962, Traffic Classification through Joint Distributions of Packet-Level Statistics**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Future work will be focused on introducing more raw feature sequences to improve the accuracy. | . |  |

## 

## **Refid: 963, User traffic classification for proxy-server based internet access control**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Our system is also practical as time consuming tasks are only performed once a month while daily tasks can be performed swiftly. Our results may be further improved by trying other models for classification. | . |  |

## 

## **Refid: 973, Detecting and diagnosing anomalies in cellular networks using Random Neural Networks**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| As future work, we shall dig deeper into feature selection approaches to improve the performance of the model with the lowest intensity anomalies of type E2. | . |  |

## 

## **Refid: 974, Skype Traffic Classification Using Cost Sensitive Algorithms**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| For future work in this area, more experiments and analysis in using cost sensitive algorithm are to be conducted in order to enhance our approach performance. Also, the proposed framework is to be tested on detecting untrained version of Malware to support zero day detection [30] against obfuscated binaries to prevent new threats from new variants. | . |  |

## 

## **Refid: 978, Intelligent IoT Traffic Classification Using Novel Search Strategy for Fast-Based-Correlation Feature Selection in Industrial**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In our future work we will check if this new method of features selection can be used to improve some other typical parameters in IoT networks, like energy consumption of routing decisions. | . |  |

## 

## **Refid: 980, Certificate-aware encrypted traffic classification using Second-Order Markov Chain**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In future work, we plan to further investigate other features in encrypted traffic to dig inspiration and improve the classification accuracy. | . |  |

## 

## **Refid: 981, Accurate classification of P2P traffic by clustering flows**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Because of its simplification and flexibility, we believe that our approach provides a promising reference for P2P traffic classification in high-speed network. More types of P2P applications are planned to address in the future. | . |  |

## 

## **Refid: 984, Privacy-Preserving Internet Traffic Publication**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  | . |  |

## 

## **Refid: 985, A multi-variate classification approach for the detection of illicit traffic**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  | traffic encryption is being widely deployed for the protection of on-line communications and most of the IDSes and anti-virus do not cope with the limitations imposed by this technology. In addition, several technical, physical and legal restrictions prevent a deep analysis of the network traffic |  |

## **Refid: 992, Mining Mobile Internet Packets for Malware Detection**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In the future, the framework can be extended. Some new mining algorithms can be applied on this framework. Mobile device security should continue to be approached through large-scale data of mobile internet | The traffic volume of mobile internet is extremely large. It is a great challenge to discover useful knowledge from it |  |

## 

## **Refid: 997, Impact of Asymmetric Routing on Statistical Traffic Classification**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Finally, a few words on future activities. We see this work continuing in several directions. First, we are studying how the findings detailed in this paper hold when analyzing other data sets, especially recent ones captured on core ISPs. Second, we plan to extend the analysis to UDP traffic in order to determine if the statistical correlation found for TCP flows holds, for example, for P2P traffic. Finally, with the practice of hot-potato routing becoming more and more standardized among Tier-1 networks, one might wonder if there is an even lower “information bound” that traffic classifiers will need to face in the future. In fact, not only they might need to be called to work on half-flows, but even such directional information could be incomplete, because of per-packet routing decisions, which might divert some of the packets composing each halfflow away from the classification device. We plan to study those issues in a future work. |  |  |

## 

## **Refid: 999, An investigation on identifying SSL traffic**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| There are several areas for future work is given that this is only investigative in nature. To this end, exploring alternative ML algorithms and other data sets are the next natural directions. Furthermore with the changes implemented between IPv4 and IPv6, it would be interesting to see how much (if at all) the results achieved in this work may be affected. Further investigation on parameter sensitivity and the affect they have on the ML algorithms may offer more explanation as to why the classifiers acted the way they did. |  |  |

## 

## **Refid: 1001, A novel approach for internet traffic classification based on multi-objective evolutionary fuzzy classifiers**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Future works will cover a more detailed analysis of more recent application-layer protocols, considering a higher number of networks. Moreover, we will also analyze and study approaches for describing the network flows in terms alternative features. |  |  |

## 

## **Refid: 1001, A novel approach for internet traffic classification based on multi-objective evolutionary fuzzy classifiers**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 1005, MINETRAC: Mining flows for unsupervised analysis amp; semi-supervised classification**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 1012, Application feature extraction by using both dynamic binary tracking and statistical learning**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 1018, On the use of Sub-Space Clustering amp; Evidence Accumulation for traffic analysis amp; classification**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 1032, Unusual internet traffic detection at network edge**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 1056, Classification of unknown mobile web traffic based on correlation coefficient measurement**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 1062, BitCoding: Protocol Type Agnostic Robust Bit Level Signatures for Traffic Classification**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 1065, On the Impact of Packet Inter Arrival Time for Early Stage Traffic Identification**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## .

## **Refid: 1068, ProDigger: Towards Robust Automatic Network Protocol Fingerprint Learning via Byte Embedding**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 1071, Machine learning based encrypted traffic classification: Identifying SSH and Skype**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Future work will follow similar lines to compare the classification based approach against other clustering based approaches and to generate more data sets to test the robustness of the classifier for the classification of other encrypted applications, such as SSL and virtual private network tunnels. Moreover, the application of this approach to encryption algorithm identification will also be explored. |  |  |

## 

## **Refid: 1077, A cascade forest approach to application classification of mobile traces**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Future work will follow similar lines to compare the classification based approach against other clustering based approaches and to generate more data sets to test the robustness of the classifier for the classification of other encrypted applications, such as SSL and virtual private network tunnels. Moreover, the application of this approach to encryption algorithm identification will also be explored. |  |  |

## 

## **Refid: 1077, A cascade forest approach to application classification of mobile traces**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In the future, we will do some theoretical analysis on the proposed scheme and evaluate its performance using multiple real-world datasets. |  |  |

## 

## **Refid: 1080, Internet traffic data categorization using particle of swarm optimization algorithm**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| The PSO algorithm takes more time for the selection of estimated value of constraints. The values of constraints influence the cluster quality during process of data. In future used optimization technique for selfselection of optimal cluster for internet traffic data. |  |  |

## 

## **Refid: 1084, Hashdoop: A MapReduce framework for network anomaly detection**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| However, we observed that dividing traffic into too many splits may cause adverse results because each split may contain insufficient traffic for the statistical tests of anomaly detectors. This tradeoff is dependent on the anomaly detector used and will be studied in future work. The implemented proof of concept benefits from Hadoop scalability, scheduling, and fault tolerance, which are particularly useful for anomaly detection. In future work, we are planning to investigate these benefits in order to run different detectors in parallel and combine them as it is done in MAWILab [7] |  |  |

## 

## **Refid: 1086, Accelerating Decision Tree Based Traffic Classification on FPGA and Multicore Platforms**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| It will be interesting to explore heterogeneous platforms for traffic classification. For example, the Zynq-7000 All Programmable System-on-Chip (APSoC) [40] is an attractive platform; on such a platform both software and hardware engines can be deployed. |  |  |

## 

## **Refid: 1091, High-throughput hash-based online traffic classification engines on FPGA**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In the future, we plan to apply the hashing modules developed to other problems such as packet classification. The conversion technique presented in this paper can also be generalized for other generic classification problems. For our design, a thorough comparison among various platforms is also of great interest. |  |  |

## 

## **Refid: 1093, Simple CART based real-time traffic classification engine on FPGAs**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In the future work, we plan to optimize the bitmap sizes to provide more memory saving and modify the proposed algorithm while also considering the feature set clustering to further improve the accuracy metric. |  |  |

## 

## **Refid: 1094, Automatically mining application signatures for lightweight deep packet inspection**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In future work, we will explore ways to use the combination of our generated payload-based signatures and flow statistics (i.e., packet inter-arrival time, packet size, flow duration) to further enhance the classification accuracy for more applications. In addition, the Stream Control Transmission Protocol (SCTP) attracts more attention in recent years since it has many advantages in terms of both security and reliability [22]. We will further extend this work to classify the applications over SCTP in the future work. |  |  |

## 

## **Refid: 1095, Sparse-Clustered Network with Selective Decoding for Internet Traffic Classification**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Our future work will focus on improving accuracy by artificially increasing the number of activated neurons in the output cluster in order to maximize flow differentiability |  |  |

## 

## **Refid: 1096, Traffic Classification Based on Zero-Length Packets**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  | Further note that few applications such as IPsec and UPnP have very few instances to attain rigorous conclusions. Note that there are 19 labels that our classifier could not detect at all. Nonetheless, most of these labels had only a single aAPDU exchange. This essentially reveals the limitations of our a-APDU based classification. Small number of a-APDU exchange indicates on a half-duplex communication pattern (such as keep-alive messages or one way notifications). This can be seen for example in dropbox, POP3, IMAPS, etc. Such a half-duplex pattern is not long or detailed enough for our classifier. Moreover, we assume that applications which significantly change their behavior based on the platform on which they run may also mislead the classifier |  |

## 

## **Refid: 1098, High-throughput traffic classification on multi-core processors**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| to explore more optimization techniques to further enhance the performance of the traffic classifiers. |  |  |

## 

## **Refid: 1099, Large traffic flows classification method**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Compared with FSM, searching algorithm based on IGR can locate the partition threshold when separating large and small flows according to minimizing the data complexity of Slarge. A data complexity metric is devised by analyzing boundary complex on Slarge. The threshold is in generalization comparing with FSM because it is independent of classification performance on validation datasets. Byte accuracy averagely increased by 10.5% comparing with FSM. However, this paper only evaluates the performance of our approach on classifying traffic flows. The research on QoE-driven network management issues could be done further, e.g. designing experiments to evaluate traffic management based on metrics of QoE. |  |  |

## 

## **Refid: 1104, Optimal supervised feature extraction in internet traffic classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 1107, Machine learning based internet traffic recognition with statistical approach**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| There is still scope of further improvement in accuracy and reduction in training time to great extent. These are very important aspects of the real time internet traffic classification process. |  |  |

## 

## **Refid: 1110, Internet traffic classification using machine learning approach: Datasets validation issues**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 1110, Internet traffic classification using machine learning approach: Datasets validation issues**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In future work, we plan to choose a combination of features and to make classifiers on the basis of feature space. We will also consider the score level fusion algorithm on the basis of the score map. We will also conduct a largescale experiment (for example, including more applications) to demonstrate the effectiveness of using score level fusion |  |  |

## 

## **Refid: 1112, Lightweight Internet Traffic Classification: A Subject-Based Solution with Word Embeddings**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| The encouraging results already achieved are stimulating further related research activities. In particular, we are investigating and experimenting more complex logic for word extraction from domain names as well as featuring stop words that are not taken into account during the classification, in order to improve the already satisfying classification performance indicators. In addition, our solution may be easily combined with other classifiers since it features a similarity score given by cosine similarity. |  |  |

## 

## **Refid: 1118, IP traffic classification in NFV: A benchmarking of supervised Machine Learning algorithms**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| As future work, we intend to generate and collect traffic data in a wider environment, for instance, by using a private cloud based on OpenStack aiming to with more data and types of services. In this way, we can verify the behavior of supervised algorithms in larger networks. |  |  |

## 

## **Refid: 1120, Low complexity, high performance neuro-fuzzy system for Internet traffic flows early classification**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Future works will be focused on system design and implementation on an Altera FPGA board, pushing on sustainability of higher link rates |  |  |

## 

## **Refid: 1151, Analysis of VBR coded VoIP for traffic classification**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 1157, Application traffic classification using statistic signature**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| However, there are several problems including the processing of unclassified traffic. In a future work, we plan to research a method that can classify Internet traffic with high accuracy and completeness using statistic signature. We also plan to research the various factors that can influence the traffic classification method using statistical information. |  |  |

## **Refid: 1160, Early Stage Internet Traffic Identification Using Data Gravitation Based Classification**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 1166, Detection of SIP signaling attacks using two-tier fine grained model for VoIP**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 1180, A Framework for QoS-aware Traffic Classification Using Semi-supervised Machine Learning in SDNs**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 1184, Internet traffic classification based on Min-Max Ensemble Feature Selection**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 1184, Internet traffic classification based on Min-Max Ensemble Feature Selection**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| For future work, we plan to find better feature in machine learning for Internet traffic classification, and maybe Tsallis entropy is a good choice, yet we have not try it. We would also improve our on-line system for better CPU and memory usage using data streaming algorithms for estimating entropy [27]. |  |  |

## 

## **Refid: 1186, Traffic classification using probabilistic neural networks**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 1190, Pushing intelligence to the network edge**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| As future work, we plan to consider the abnormal traffic detection algorithm, the security rules configuration, as well as the implementation of the traffic analysis module into an SDN based controller |  |  |

## 

## **Refid: 1194, Cooperative learning for online in-network performance monitoring**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 1195, Online hybrid internet traffic classification algorithm based on signature statistical and port methods to identify internet**

## 

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 1198, Minimal dataset for Network Intrusion Detection Systems via dimensionality reduction**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  | Requirement for more input parameters from user (render the automation difficult), • Longer time to complete dimensionality reduction on KDD dataset (with some lasting more than 10 minutes). This is unacceptable for intrusion detection as we need a near real-time |  |

## 

## **Refid: 1204, A New Feature Selection Method for Internet Traffic Classification Using ML**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| As a future work, we intend to improve our feature selection method on reducing the computation overhead. |  |  |

## 

## **Refid: 1205, Abacus: Accurate behavioral classification of P2P-TV traffic**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| As part of our future research, we aim at digging this issue further, by observing the same traffic from multiple vantage points in the network, and applying the classification to different subsets to verify to what extent the above considerations hold in practice |  |  |

## 

## **Refid: 1217, Efficient application identification and the temporal and spatial stability of classification schema**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 1219, Application-based feature selection for Internet traffic classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| We consider a number of future extensions to this work. We intend to carry a systematic study of selective features for key applications like BitTorrent or HTTP streaming with our method. Also, we have considered TCP traffic only so far. However with the growing trend of UDP traffic, we would like to generalize the method to handle UDP traffic as well |  |  |

## **Refid: 1225, Exploiting unlabeled data to improve peer-to-peer traffic classification using incremental tri-training method**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| As one of our future work, we will focus on reducing the computation cost, and also, explore the incremental semi-learning algorithms which do not use windowing techniques |  |  |

## 

## **Refid: 1226, Histogram-based traffic anomaly detection**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| An important issue remains open, which is the reduction, or ideally elimination, of false positives. False positives are an inherent problem with all anomaly detection systems, as both normal and abnormal conditions can occasionally result in the same observable characteristics. In our work, we found a medium number of false positives. The successful and widespread use of feature-based anomaly detection systems requires lasting efforts for reducing the occurrences of costly incorrect alarms. A number of promising paths exist that we plan to explore in our future research |  |  |

## 

## **Refid: 1227, Classification of traffic flows into QoS classes by unsupervised learning and KNN clustering**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 1227, Classification of traffic flows into QoS classes by unsupervised learning and KNN clustering**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| The way our model represents a flow has a few limitations and it can evolve, with further work, in a more accurate protocol description. For example, it needs to deal with out-of-order packets, packet loss, and fragmentation in a robust way. We plan to address these and other open issues in a future work. Finally, we are currently working on a systematic comparison of this SVM-based classifier with other approaches, including those based on Gaussian Mixture Models (GMM), neural networks and payload analysis. |  |  |

## 

## **Refid: 1262, Intelligent IP traffic/flow classification system**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Feasibility of using the feature in real-time classification scenario, where the classification has to be done in a specific period of time, will be invested in our future work |  |  |

## 

## **Refid: 1265, K-dimensional trees for continuous traffic classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| We plan as future work to study the alternative solution of incrementally updating the old model with the new information, instead of creating a new model from scratch. In addition, it is possible to automatically update the list of relevant ports by using the training data as a reference  As future work, we plan to perform a more extensive performance evaluation of our continuous training system with long-term executions in order to show the large advantages of maintaining the classification method continuously updated without requiring human supervision. |  |  |

## 

## **Refid: 1277, An automatic application signature construction system for unknown traffic**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  | In this work we use the implicit form of application signatures, which are trained with byte feature vectors representing the initial payload content of each traffic flow. Although these signatures are excellent for classification regarding accuracy and efficiency, one can hardly learn any knowledge from them. This may become a practical problem when a network administrator wants to know more about the newly discovered applications. However, other techniques can be exploited to fulfill this task. For instance, we can derive common substrings from each cluster, which can serve as keywords of the unknown applications. Another related issue comes from the mapping of multiple clusters to one application. As discussed before, this feature can provide a fine-grained view into an application and let us learn more knowledge from its diversely distributed characteristics. However, it is easy to get confused when the administrator is suddenly informed that there are so many classes of new applications emerging in the network. Merging the clusters is a possible solution, which is to be explored in future work. |  |

## 

## **Refid: 1307, Monitoring, analysis, and filtering system for purifying network traffic of known and unknown malicious content**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Future work will include development and evaluation of additional plug-ins (including dynamic/behavioral plugins) and advanced risk-weighing algorithms. We plan to explore the Signature Builder method further and to evaluate additional methods for detecting and extracting functions in binary code and additional methods for ranking and selecting the best signature out of the collection of candidates. We also plan to evaluate the system for various NSP topologies and for various configurations of cleaning appliances and monitoring locations. Moreover, polymorphic malware that is currently not treated by eDare requires future research leveraging from a collaboration between eDare’s NMDM and the eDare Agents |  |  |

## 

## **Refid: 1310, A semi-supervised clustering method for P2P traffic classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| With more P2P metrics studied, it would be developed to classify P2P traffic at connection level in the future work. |  |  |

## 

## **Refid: 1311, Internet traffic classification based on fuzzy kernel K-means clustering**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Experimental results show that this method has higher cluster accuracy to provide QoS guarantees according to all kinds of Internet application. Moreover, several opportunities exist for future work. We need more experiments to find out which flow features are suitable for improving traffic classification accuracy. |  |  |

## 

## **Refid: 1312, Analysis of the impact of sampling on NetFlow traffic classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In the paper, we also identified several limitations, common to most machine learning techniques, which constitute an important part of our future work. Nevertheless, the main limitation in this research area is the lack of publicly available traces to be used as a common reference for researchers working on this topic. In this direction, an additional contribution of this paper is that we have opened our traffic traces to the research community. These traces were collected at a large university network and cover a wide range of days and hours |  |  |

## 

## **Refid: 1314, Online wireless mesh network traffic classification using machine learning**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| This method is also suitable for online network traffic classification. Our experimental results so far are promising for this research direction in wireless mesh network, and several opportunities exist for future work. Supervised machine learning method is the requirement on a large number of labeled training samples. We can propose semi-supervised classification method to solve this issue for wireless mesh network traffic classification. |  |  |

## 

## **Refid: 1316, Online internet traffic classification based on proximal SVM**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Several opportunities exist for future work. Supervised machine learning method is the requirement on a large number of labeled training samples. We can propose semi-supervised classification method to solve this issue. Moreover, we also need more experiments to find out which features are efficient and suitable for improving the classification accuracy |  |  |

## 

## **Refid: 1317, Real-time encrypted traffic identification using machine learning**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| The proposed real-time encrypted traffic identification methodology can identify encrypted Skype applications using flow character-based approach with high accuracy and low overheads. Furthermore, we need more experiments to find out which features are suitable for improving the encrypted traffic identification accuracy. |  |  |

## 

## **Refid: 1330, Early identification of peer-to-peer traffic**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| On the other hand, our method is able to classify only a few traffic types, leaving open the question if it can scale up to handle many more protocol types, too. Preliminary results to handle unknown traffic were also provided, but further steps are necessary in this direction. Moreover, the performed tests were made on actively collected data: in this case the ground truth is known without doubt, but the traffic patterns are less realistic than in real traces. However, we believe that the fact that our algorithms use only the first 16 bytes of the flows makes our method less sensitive to such errors (apart from the fact that the TP and FP rates are highly dependent on the composition of the whole data set). Nevertheless, measurements on real traffic traces (with high quality labeling) should be performed in the future. |  |  |

## 

## **Refid: 1336, High-performance intrusion detection using OptiGrid clustering and grid-based labelling**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| The better extraction procedure of training data and the reduction of parameters are open issues for the future work. Research of IDS is the spiral of fight against attackers. In the middle of the spiral, we hope our research becomes contribution to the fight. This work was partially supported b |  |  |

## 

## **Refid: 1358, Feed-forward neural networks for accurate and robust traffic classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| This research is a valuable area for future work. Because of continuous development of new internet applications, combining neural networks with unsupervised clustering techniques to automatically identify and classify new traffic classes is also a promising direction of future research. |  |  |

## 

## **Refid: 1367, Feature selection for optimizing traffic classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In our experiments, the TPR of each application achieved with the robust WSU\_AUC-selected features was not consistently superior. But this case is rare. For instance, only classifying FD and INT with NBK classifier, the SU\_ACC-selected features were more efficient. On the other hand, the TPRs of identifying the minority classes with the robust WSU\_AUC-selected features look better, but far from perfect. For example, the TPR of identifying ssh application is not high on the CAIDA data sets. The further research on improving the specific application, such as ssh, is left for future work. The characteristics of feature selection algorithms can be described as search organization, generation of successors and evaluation measure [36]. In the WSU\_AUC algorithm, we only consider optimizing the evaluation measure. Search organization and generation of successors dominate the computation overhead of feature selection algorithm. In the future, we will improve the performance of WSU\_AUC by means of optimizing search algorithms. Besides, we will study deeper how to detect the underlying changes caused by dynamic data flows and correspondingly how to rebuild the model to adapt these changes |  |  |

## 

## **Refid: 1372, Machine learning-based classification of encrypted internet traffic**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 1397, Traffic classification model based on integration of multiple classifiers**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Based on this research, the next step is mainly: based on this flow data and the measure of flow properties proposed, to provide data to support the further study, is also able to improve the identification of multi-classifier model to better meet the online traffic identification |  |  |

## 

## **Refid: 1402, SSPC algorithm based on three different methods for online Skype traffic classification**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 1404, Encrypted Traffic Classification Based on an Improved Clustering Algorithm**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 1407, Out-of-sequence traffic classification based on improved dynamic time warping**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Speed up IDTW calculation |  |  |

## .

## **Refid: 1408, Towards fingerprinting malicious traffic**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Our future works fall into classify the malicious traffic accordingly to malware types and families, and deploying the model on a network in order to test its performance on realtime traffic. |  |  |

## 

## **Refid: 1413, Automated Dataset Generation for Training Peer-to-Peer Machine Learning Classifiers**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In future works, we plan to analyse different policies for updating the on-line ML classifier and also implement the on-line part of our proposed system (on-line ML classifier) on hardware in order to classify P2P traffic and evaluate TSTDG in an on-line network. In addition, our system can be integrated with a network control system in order to classify and control P2P traffic |  |  |

## 

## **Refid: 1414, Retraining mechanism for on-line peer-to-peer traffic classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 1418, Classification Research on SSL Encrypted Application**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| The next phase of our work should include the following aspects: First, it is necessary to use C4.5 method in the actual network environment. Second, we should be able to find new methods and features to identify more encrypted applications based on SSL protocol. |  |  |

## 

## **RRefid: 1420, Automatically mining application signatures for lightweight deep packet inspection**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In future work, we will explore ways to use the combination of our generated payload-based signatures and flow statistics (i.e., packet inter-arrival time, packet size, flow duration) to further enhance the classification accuracy for more applications. In addition, the Stream Control Transmission Protocol (SCTP) attracts more attention in recent years since it has many advantages in terms of both security and reliability [22]. We will further extend this work to classify the applications over SCTP in the future work. |  |  |

## 

## **Refid: 1423, Application traffic classification at the early stage by characterizing application rounds**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| future directions: constructing sharable traffic traces with reliable ground truth; developing the scalable traffic classifiers to support the high-speed classification with the increasing speed of network links; classifying the encapsulated, encrypted, and multichannel application flows; combining the different classifiers to improve the accuracy and performance and decrease the cost in a multi-classifier platform; and publishing the open source traffic classifier to promote collaboration in reliable evaluation |  |  |

## 

## **Refid: 1428, Toward an efficient and scalable feature selection approach for internet traffic classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Future work will further reduce such computational requirements by parallelizing the proposed approaches in distributed systems.  In the future, we will work on developing a new approach to address the sensitivity of the baselines methods and the LOA approach to variations in the traffic data sets. |  |  |

## 

## **Refid: 1442, Classifying peer-to-peer applications using imbalanced concept-adapting very fast decision tree on IP data stream**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 1449, Traffic classification model based on fusion of multiple classifiers with flow preference**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 1453, Zero-day traffic identification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| With system update, the proposed scheme can further achieve more fine-grained classification of zero-day traffic. Moreover, a quantitative analysis on flow correlation was also provided confirm the effectiveness of the proposed scheme |  |  |

## 

## **Refid: 1470, A comparison of improving multi-class imbalance for internet traffic classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Our next work would be some insights as below. Address influence of multi-class imbalance to Internet traffic classification through using new flow features. Combine sampling with cost-sensitive learning together to carry out traffic classification, e.g. sampling is used to improve class imbalance first then the cost-sensitive learning is used to generate classifier for Internet traffic classification. We would point, however, it is important to learn what the class distribution of the final training data should be first. Then it might be beneficial to discard some of the majority samples because a traffic flow dataset is enormous usually in order to reduce the training set to the desirable size and then employ a cost-sensitive learning algorithm again. If the cost information is not known, a measure such as the area under the ROC curve could be used to measure classifier performance and then the proper cost ratio/class distribution is determined. |  |  |

## 

## **Refid: 1471, Internet traffic clustering with side information**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In the future work, we plan to explore the automatic estimation of the best value for cluster number with considerations of the side information |  |  |

## 

## **Refid: 1490, Selection of on-line features for peer-to-peer network traffic classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 1495, Feature evaluation for early stage internet traffic identification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Although the combined feature set does not outperform the other feature sets significantly, we infer that more effective feature selection methods can pick out combined feature sets that are more effective than single type of features. And this is an important future work of us. |  |  |

## 

## **Refid: 1498, High-performance traffic classification on GPU**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In the future, we plan to develop high-speed feature extraction algorithm with throughput at par with the traffic classifier in this paper, so that the entire traffic classification process chain could be pipelined. We observe that when the number of features increases, the throughput decreases in both high-throughput and low-latency implementations. In the high-throughput implementation, the deterioration is because of more computation per thread; whereas in the low-latency implementation, the deterioration is due to significant amount of time spent for synchronization among threads. In the future, we plan to develop an algorithm that makes the best of both designs to get better performance for more number of features |  |  |

## 

## **Refid: 1506, Hybrid P2P traffic classification with heuristic rules and machine learning**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  | Currently, there are still some limitations in our proposed scheme. Although the first-step classifier can classify specific P2P traffic well, the second-step classifier still can only work in classifying P2P and non-P2P traffic. There are plans to implement a flow-level classifier that can classify specific P2P traffic generated by different P2P applications. In addition, more pattern heuristics have to be created and validated to reduce the false-positive rate in the flow-level classifier. Our proposed scheme does not support the detection of UDP traffic of P2P. The traffic dataset will be extended to include the UDP traffic of P2P |  |

## 

## **Refid: 1534, Automatic traffic classification using machine learning algorithm for policy-based routing in UMTS–WLAN interworking**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| This work can be extended to increase the accuracy of the traffic identification with suitable unsupervised learning algorithm which exhibits minimum classification time and reduced manual intervention in classification. |  |  |

## 

## **Refid: 1540, Effective packet number for early stage internet traffic identification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 1554, Anomaly detection based on efficient Euclidean projection**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 1555, Statistical user behavior detection and QoE evaluation for thin client services**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Future work will focus on improving our classification accuracy, and further analyzing additional real world scenarios, such as those focused on use of thin client services in an enterprise/business setting. A further focus of this work has been on combining the potential to identify user behavior when using RDC, with the ability to make effective QoE estimations for different behavior categories under various network conditions. Our work has built on previous studies that have explored QoE for RDC services [21, 4, 6], and addressed how behavior detection can be used together with functions modeling QoE in terms of network delay, bandwidth, and loss for different types of remote desktop services. The motivation for such studies lies in the fact that such QoE estimations may then be further used for QoE-driven management mechanisms such as resource planning and dynamic resource (re)allocation in a cloud environment, or service adaptation in order to assure a satisfied customer base. An example of recent work focused on exploiting RDC application identification using statistical mechanisms for the purpose of QoE-driven scheduling is discussed in [1] |  |  |

## 

## **Refid: 1557, Robust representation for domain adaptation in network security**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| There are several remaining challenges in the domain adaptation for network security. With constantly evolving malware, conditional shift might still occur even when the new malware families are represented as outlined in this paper. There are other types of malware, some of which have not been identified or fully understood, that have different behavioral patterns making it impossible to transfer knowledge from the source to the target domain. Some of these challenges might be solved by introducing nonlinearity to the malware similarity measure. As in the presented online similarity learning, the measure could use the known samples to learn the differences between malicious and legitimate traffic. This is the direction of our future research. |  |  |

## 

## **Refid: 1561, Anomaly detection in traffic using L1-norm minimization extreme learning machine**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In the field of traffic classification and anomaly detection research, the classification is very resource-intensive, thus, we plan to make effort on selecting sparse and robust features before classifications to get rid of the most resource requirements. |  |  |

## 

## **Refid: 1564, An Autonomic Traffic Classification System for Network Operation and Management**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Although this work is mostly completed and our system has been already deployed in a production network (CESCA), there are three lines of future work we plan to study. First, the increasing importance of Content Delivery Networks has decremented the power of IP-based classification techniques. Consequently, it would be interesting to study the inclusion of more classification techniques as those based on host-behaviors [4, 19, 20]. Second, we plan to include new techniques [33– 35] to reduce the amount of unknown traffic and improve the generation of the ground truth. Finally, we plan to study the viability of using stream-oriented ML techniques, given that its streaming operation seems more suitable for processing the network traffic. |  |  |

## 

## **Refid: 1571, MTRAC - Discovering M2M devices in cellular networks from coarse-grained measurements**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In the future, we plan to extend the MTRAC system to work with longer periods of data. As it already appears, in the last three weeks of Fig. 5b the FPR gradually increases. Therefore, we plan to apply advanced on-line learning methods to cope with this concept drift. In addition, we want to study the increased FPR over weekends in more detail. During weekends, the FPR is up to approx. 10% worse than on weekdays. Last but not least, we are planning to extend our approach from using session based to flow based data. Since the number of flows is a lot higher than the number of sessions, more information is available per device which might improve the performance of the overall approach. . |  |  |

## 

## **Refid: 1575, Designing an Internet Traffic Predictive Model by Applying a Signal Processing Method**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| in future works, we plan to identify more informative features. To increase the accuracy and stability for detecting abnormal network behaviors in a different network environment, we are going to utilize more network features, such as IP addresses, protocols (including port information), network services (e.g. HTTP, Telnet, SSH, or else.), and TCP flags. Although numerous network detection techniques have been proposed, it is still difficult to determine specific attack types. We plan to extend our research to detect attack types including DDoS (distributed denial-of-service), Buffer overflow attack, Surveillance sweep, or others. |  |  |

## 

## **Refid: 1576, How Robust Can a Machine Learning Approach Be for Classifying Encrypted VoIP?**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  | No signature-based method in traffic classification can be perfect, given that there can always be some new (unseen) applications. Thus, the major challenge for traffic Fig. 14 Sensitivity of C5.0 FPR for changing the confidence factor parameter Fig. 15 Sensitivity of the number of C5.0 rules for changing the confidence factor parameter 862 J Netw Syst Manage (2015) 23:830–869 123 classification in general is evasion. All classification methods can be evaded. For example, a payload-based approach can be evaded by encrypting the packet payload and a port-based approach can be evaded by changing the port numbers dynamically. However, approaches based on flow statistics using packet size and inter-arrival time attributes are sensitive theoretically to altering these attributes. If attackers want to evade the proposed method they can modify the size of the packets in the entire connection by padding the packet payloads randomly. The accuracy of the proposed method might be decreased if those features that depend on packet size were modified from the application behaviour. However, it is not that easy to obfuscate application behaviour without presenting a large amount of overhead or changing the quality of the application. For example in the VoIP case, the amount of padding that can be implemented also has its limit. Too much padding may corrupt Rule 215: (879366/168737, lift 1.1) min\_fpktl <= 139 min\_bpktl > 48 -> class NOTSKYPE [0.808] Rule 216: (9456/3, lift 3.5) std\_fpktl <= 0 max\_bpktl <= 321 std\_bpktl > 80 max\_fiat <= 26620 mean\_biat > 34983 mean\_biat <= 330261 -> class SKYPE [1.000] Rule 217: (1801, lift 3.5) min\_fpktl > 131 duration <= 126901 total\_fvolume <= 132 total\_bvolume <= 63 -> class SKYPE [0.999] Rule 218: (798, lift 3.5) min\_fiat > 0 min\_fiat <= 3 -> class SKYPE [0.999] Rule 219: (6044/3, lift 3.5) min\_fpktl <= 94 std\_fpktl <= 2 min\_bpktl > 61 min\_bpktl <= 79 max\_bpktl > 244 max\_bpktl <= 298 std\_bpktl > 103 -> class SKYPE [0.999] Rule 220: (782, lift 3.5) min\_fpktl > 34 min\_fpktl <= 50 std\_fpktl > 2 min\_bpktl > 46 min\_bpktl <= 79 mean\_bpktl > 164 max\_bpktl <= 595 proto > 6 total\_fpackets <= 8 total\_bvolume > 359 -> class SKYPE [0.999] Fig. 16 An example of a C5.0 solution for Skype detection based on the flow feature set J Netw Syst Manage (2015) 23:830–869 863 123 the voice service to the extent that the parties talking can no longer hear or understand each other. In this case, whether the application can evade the classifier or not is irrelevant given that it is no longer a useful application in terms of carrying voice data over the IP. Another limitation of any classification system is obtaining (generating) the training data set. The generality and accuracy of the classifier depends on the quality of the training data sets. A meaningful and representative training data set is hard to find and generating one is resource and time consuming. Moreover, since the classifier generates the signatures automatically from the training data set, the accuracy of the classifier might decrease if the signatures/models from the trained classifiers are applied to network traffic that have different characteristics or behaviour (such as new applications that are developed or old applications that change their behaviour). Indeed, in such cases, the signatures/models need to be updated by retraining the classifiers. That is why it is very important to conduct robustness analysis on such classifiers. For instance, in this research, C5.0 signatures have the best consistent performance in the robustness criteria and the signatures can classify Skype P2P VoIP traffic in a trace robustly if the characteristic of the flow features in the trace fall within the range |  |

## 

## **Refid: 1585, On Internet Traffic Classification: A Two-Phased Machine Learning Approach**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Furthermore, the substantive accuracy of the present approach in achieving highly granular per-flow application identification and the computational efficiency in comparison with other machine learning classification methodologies paves way for future work in extending this method to include other applications for real-time or near real-time flow based classification. |  |  |

## 

## **Refid: 1590, Modelling the multi-layer artificial neural network for internet traffic forecasting: The model selection design issues**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  | Despite that the study did not make any attempt to determine an optimal values for the various factors considered, it has shown that careful experimentation is required to choose appropriate values for each of the design issues. Therefore, the multilayer perceptron should not be applied blindly to Internet traffic forecasting. |  |

## 

## **Refid: 1591, Spot the hotspot: Wi-Fi hotspot classification from internet traffic**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In future we intend to expand and diversify the dataset in terms of demographics and geography. Furthermore, we aim to classify the hotspot type using other information sources. For example, classify the hotspot type locally on the device using information that can obtained by applications (e.g., Wi-Fi connections, internet usage statistics and sensors). |  |  |

## 

## 

## **Refid: 1607, Effectiveness of Statistical Features for Early Stage Internet Traffic Identification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Features using for identification application should be carefully selected. How to select high effective feature sets for early stage traffic identification is an important problem to be resolved in our future work. |  |  |

## 

## **Refid: 1619, A preliminary performance evaluation of K-means, KNN and em unsupervised machine learning methods for network flow**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 1635, Classification of VoIP and non-VoIP traffic using machine learning approaches**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Our future studies will concentrate on examining other features to increase the accuracy. We will also study the accurate and detailed classification methods for the studied traffic by considering other ML techniques. |  |  |

## 

## **Refid: 1636, DBStream: A holistic approach to large-scale network traffic monitoring and analysis**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| we want to investigate the possibility of extending DBStream by replacing the database engine PostgreSQL with either the parallel database system Greenplum [48], or a MapReduce based large-scale data processing framework like, e.g., Spark [16]. Indeed, this would be a logical extension of the current single machine DBStream architecture to a cluster system, thus enabling scale-out properties found in modern big data processing frameworks. Furthermore, we have deployed DBStream in the intelligent transportation systems domain, and plan its adoption also in other application domains with similar properties such as smart grid and smart city. In fact, data from those application domains has similar properties. Data arrive as high volume data streams and the analytic questions can be addressed utilizing DBStreams CEL language. Preliminary results show that DBStream can be used to store and analyze data from those domains as successfully as from computer networks. |  |  |

## 

## **Refid: 1652, Effective Feature Selection for 5G im Applications Traffic Classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| However, there is still a gap for further research in the 5G instant messaging (IM) traffic classification. A new approach should be designed to select robust feature for IM applications traffic classification and this is our future research work. |  |  |

## 

## **Refid: 1655, Classification of network traffic using fuzzy clustering for network security**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Additional Research to test various feature subset reduction techniques is planned. Reducing the number of features will reduce the computational complexity of the FCM approach. Further research is planned to test and evaluate the membership threshold in clusters for identifying attacks to determine the optimal membership for identification of malicious packets. Once these issues have been resolved, testing using datasets than the 4GB KDD dataset and in a real world environment will be conducted. There are also plans to incorporate the ability to learn new clusters or adjust cluster centroids based upon newly discovered attack modes |  |  |

## 

## **Refid: 1657, A search for computationally efficient supervised learning algorithms of anomalous traffic**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  | However, most of these techniques have limitations due to the use of expert-based misuse detection methods and statistical-based abnormal behaviour detection methods. |  |

## 

## **Refid: 1658, Multi-stage feature selection for on-line flow peer-to-peer traffic identification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## 

## **Refid: 1661, Link-layer device type classification on encrypted wireless traffic with COTS radios**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Our investigation revealed that an average of 3300 packets and an observation window size of about 1200s to sniff them are needed to build effective signatures to achieve this accuracy in our scenarios. It is observed that this estimation can vary significantly depending on the status of the devices present in the target network, and further investigation in this direction is planned as a future work. Finally, we have tested our classifier in an open and uncontrolled university area, and we are successful in detecting devices like laptops, smartphones and desktop computers in high numbers. However, our analysis in this case shows that it requires a larger number of devices of certain types like smart lights and cameras to achieve more precise classification. |  |  |

## 

## **Refid: 1671, P2P and P2P botnet traffic classification in two stages**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  | In general, existing hybrid classification schemes require a large amount of computation and time-consuming. To overcome these limitations, several methods are tried in our proposed scheme. Connection heuristics are used in the first stage. The connection heuristics reduce the amount of packets processed that need to be processed during analysis and flow feature collection by 36.51 and 59.48 %. The flows needed for classification are greatly reduced by 43.20 % in the first stage and 59.57 % in the second stage. The first stage filters most of the non-P2P flows and accelerates the secondstage classification. Since the flow features in the second stage are collected in the first stage, we do not have to collect the flow features again in the second stage. Early classification can be applied because the flow features used by the statistics-based classifier and the P2P botnet traffic classifier are packet size-related features, which can be obtained before a flow has completed. There are still some limitations to our proposed scheme. The statistics-based classifier can be expanded to classify particular P2P traffic generated by different P2P applications. In the same way, the P2P botnet traffic classifier can also be expanded to detect specific P2P botnet traffic generated by different malwares in a P2P botnet. In addition, only TCP traffic is considered in this paper, and the traffic dataset can be extended to include UDP traffic of P2P and P2P botnet. |  |

## **Refid: 1677, Imbalanced traffic identification using an imbalanced data gravitation-based classification model**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Third, data-level imbalanced classification algorithms performed poorly with imbalanced traffic data, where SMOTEBagging and SMOTEBoost performed like random classifier patterns in the identification experiments. We also found that C4.5CS performed well in the experiments. Therefore, using a single model for imbalanced traffic identification might not be the best solution and it may be more effective to build a hybrid model to solve this problem, which is an important issue that will be addressed in our future research |  |  |

## **Refid: 1680, Optimizing an artificial immune system algorithm in support of flow-Based internet traffic classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In future research it would be very simple to implement the same optimization in a negative selection algorithm, and get the same speedups. Based on a literature review, we have not found the techniques used in this paper anywhere else in the literature of optimized AIS algorithms  . In future research we can investigate different ways to optimize the margin of each base classifier during the training portion of the algorithm by optimizing the placement of its center as well as it radius. Furthermore, the scheme used to set the radius of each antibody is also very simple, since itis an approximation ofthe class boundary and does nottake into account the generalization ability of the algorithm. It might be interesting to try other ways of finding an optimal antibody radius which allow an antibody to misclassify examples, but increase the generalization ability of the classifier. Since beginning work on this AIS-inspired algorithm we have found it to be particularly insensitive to the choice in parameters. Unlike other classification algorithms,this algorithm does not seem need to have its parameters set up perfectly to give good performance. An open question for us is: how insensitive is the algorithm to the values of the parameters given to it? Another interesting research direction would be to apply a similar bound to the negative selection algorithm, using hyper-spheres as base classifiers. Furthermore, a bound could be drawn on this and other artificial immune system algorithms that use base classifiers that are not hyper-spheres. |  |  |

## **Refid: 1683, Efficient and robust feature extraction and selection for traffic classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| The future work will be devoted to apply the proposed approach into network engineering. |  |  |

## **Refid: 1691, A novel approach for internet traffic classification based on multi-objective evolutionary fuzzy classifiers**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Future works will cover a more detailed analysis of more recent application-layer protocols, considering a higher number of networks. Moreover, we will also analyze and study approaches for describing the network flows in terms alternative features. |  |  |

## **Refid: 1706, Ensemble network traffic classification: Algorithm comparison and novel ensemble scheme proposal**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## **Refid: 1718, Distributed detection of zero-day network traffic flows**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In practice, the network traffic can change dynamically, and new zero-day traffic can emerge in the network arbitrarily. This means that the processing of the network traffic needs happen dynamically. Our distributed framework can be run periodically, using a pre-defined periodicity (a user defined parameter), in order to capture the newly emerging zero-day traffic in the network. For example, in the context of the ISP data set (that we used in our evaluation), our system can be refreshed weekly, since the network traffic data has been collected during a seven day period. An interesting question here is how this process can be done incrementally? We left this topic for future research. |  |  |

## **Refid: 1719, Traffic classification based on incremental learning method**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## **Refid: 1735, An efficient feature generation approach based on deep learning and feature selection techniques for traffic classification**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |

## **Refid: 1740, A BasisEvolution framework for network traffic anomaly detection**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| However, there are many issues left for future work. For example, the basis generation algorithm we used is based on SVD, which is not feasible for very large data scenarios. The first issue is to find robust, fast algorithm to replace SVD, such as the pursuit principal used in OMP. Another point is that the size of data segment we used in our framework is fixed and our algorithm proceeds in batches. A more realistic algorithm should be able to tackle arbitrary new segments of data as they arrive. We plan to extend BasisEvolution in the future to tackle these issues. |  |  |

## **Refid: 1741, Tor anonymous traffic identification based on gravitational clustering**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| In the future, we plan to introduce new features and explore traffic identification method from other anonymous communication tools. |  |  |

## **Refid: 1742, Detecting malware-infected devices using the HTTP header patterns**

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
| Verification of our approach using malware samples of other platforms is left for future study. | BotDetector makes use of the HTTP header fields; hence, the system cannot react when traffic is encrypted, i.e., when communication is made with HTTPS. In addition, although the HTTP protocol is one of the most popular protocols used by malware, the current system cannot detect malwareinfected traffic based on other protocols. For instance, UDPbased protocols may not be able to be captured with our approach. We note that the scope of our work does not cover certain types of malware samples such as those used for spear-phishing email attack. As we have shown, our approach successfully detected malware samples uch as Adware, Trojan, Worm, Downloader, Ransomware, etc., which all make use of HTTP as a means of communication. We believe our approach works in a wide range as HTTP is a common way of communication widely used for various malware families. Another limitation of BotDetector is that malware developers can change the HTTP headers to evade detection; i.e., the traffic originated from malware can mimic the traffic originated from a browser. Although the case has not been major so far, such evasion could become standard in the future, at which point, we need to change the feature extraction and classification model. We also note that our targets were limited to Windows malware. As the analysis of HTTP headers is independent from the system architecture, our approach should work for malware of other platforms. Verification of our approach using malware samples of other platforms is left for future study. Despite of these limitations, we believe that the fundamental idea behind this work – finding useful features automatically – remains beneficial to discover invariants that could be used to detect malicious activities. |  |

## 

## **Refid: 1745, A scalable distributed machine learning approach for attack detection in edge computing environments**

## 

|  |  |  |
| --- | --- | --- |
| Future Direction | Limitations | Challenges |
|  |  |  |