An overview of flow-based anomaly detection

Article *in* International Journal of Communication Networks and Distributed Systems · July 2018

DOI: 10.1504/IJCNDS.2018.10014505

3 authors:

Rohini Sharma
Panjab University
16 PUBLICATIONS **59** CITATIONS
SEE PROFILE

Ajay Guleria
Indian Institute of Technology Delhi
23 PUBLICATIONS **372** CITATIONS
SEE PROFILE

Ravinder Kumar Singla
Panjab University
50 PUBLICATIONS **402** CITATIONS
SEE PROFILE

An overview of flow-based anomaly detection

Rohini Sharma*

Department of Computer Science and Applications,
Panjab University,
Chandigarh, India
Email: rohini@pu.ac.in
*Corresponding author

Ajay Guleria

Computer Centre,
Panjab University,
Chandigarh, India
Email: ag@pu.ac.in

R.K. Singla

Department of Computer Science and Applications,
Panjab University,
Chandigarh, India
Email: rksingla@pu.ac.in

Abstract: Intrusions in computer networks are handled using misuse or anomaly-based solutions. Deep packet inspection is generally incorporated in solutions for better detection and mitigation but with the growth of networks at exponential speed, it has become an expensive solution and makes real-time detection difficult. In this paper, network flows-based anomaly detection techniques are reviewed. The review starts with motivation behind using network flows and justifies why flow-based anomaly detection is the need of the hour. Flow-based datasets are also investigated and reviewed. The main focus is on techniques and methodologies used by researchers for anomaly detection in computer networks. The techniques reviewed are categorised into five classes: statistical, machine learning, clustering, frequent pattern mining and agent-based. At the end the core research problems and open challenges are discussed.

Keywords: network flows; anomaly detection; security; privacy; flow-based dataset; statistical techniques; machine learning; clustering; frequent pattern mining; software agents.

Reference to this paper should be made as follows: Sharma, R., Guleria, A. and Singla, R.K. (2018) 'An overview of flow-based anomaly detection', *Int. J. Communication Networks and Distributed Systems*, Vol. 21, No. 2, pp.220–240.

Biographical notes: Rohini Sharma holds both MCA and MTech in Computer Science and Engineering. Currently, she is pursuing her PhD from the Panjab University, Chandigarh. She is an Assistant Professor in the Department of Computer Science and Applications, Panjab University, Chandigarh, India. Her research interests include network anomaly detection and data mining.

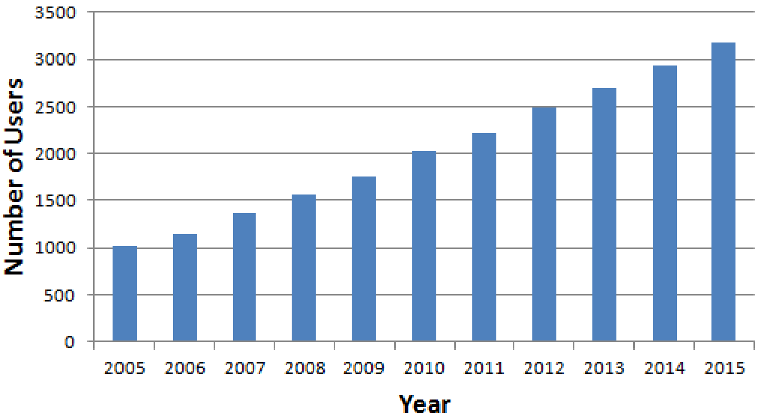
Ajay Guleria received his PhD from the National Institute of Technology, Hamirpur, Himachal Pradesh, India. He is a System Manager at the Panjab University, Chandigarh, India. His research interests include software defined networking, information centric networking, network security and vehicular ad hoc network.

R.K. Singla received his PhD in 1999 from the Panjab University, Chandigarh, India. He is a Professor at the Department of Computer Science and Applications, Panjab University, Chandigarh, India having 30 years of teaching and research experience. He has published more than 60 research papers. His current research interests are in the areas of computer and network security, open source software engineering and machine learning.

1 Introduction

Computer networks is a field which started with advanced research projects agency network (ARPANET) in 1969 and going towards internet of things (IoT). It has become an important part of our daily life which is affecting virtually everybody. Personal data and resources residing on hosts/servers are required to be secured from unauthorised use, eavesdropping, etc. from internal as well as external sources. Security is a big issue having different aspects like integrity, confidentiality, access control, etc. which can be ensured using various techniques like encryption, password, firewalls, etc. As the size of network increases, the concern for security increases. With the universal spread of internet, size of networks is increasing day by day and hence the danger of harm to resources on the system is increasing.

Figure 1 Growth of internet users (see online version for colours)



As per statistics available over www.statista.com, number of internet users in 2015 has grown to 3,174 million; the increasing trend is shown in Figure 1. Internet is a source of unwanted intrusions which may result in the halt of complete system. According to McAfee Labs Threats Report (2015) (http://www.mcafee.com/in/resources/reports/rpquarterly-threats-aug-2015.pdf), amount of malicious software (malware) is increasing with each increasing quarter as shown in Figure 2. Numerous types of malicious software's are there which helps in breaching the security of a system. Top threats of 2015 are shown in Figure 3 as given in the report by McAfee (http://www.mcafee.com/in/resources/reports/rp-quarterly-threats-aug-2015.pdf). Attacks and threats on internet are increasing in frequency and impact.

Figure 2 Increasing trend in malicious software's (see online version for colours)

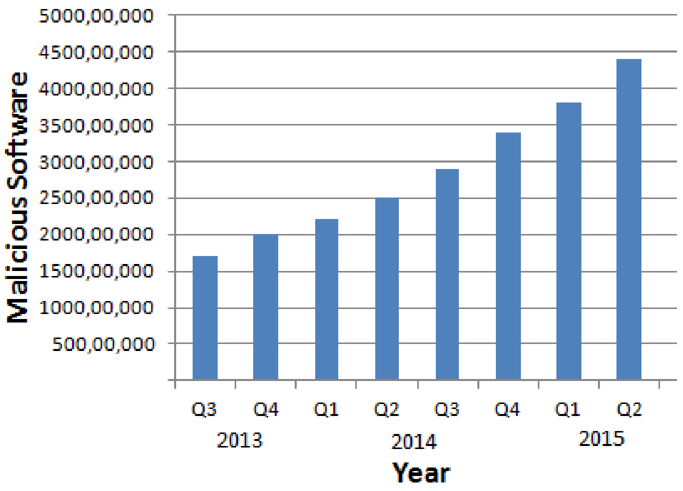
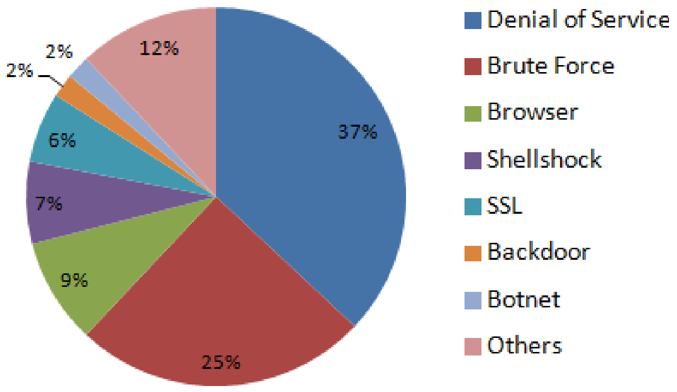


Figure 3 Top threats of 2015 (see online version for colours)



To ensure smooth running of established networks, security plays an important role. People and organisations are spending large amount of money to secure their resources like servers, systems, networks, etc. network security has been a hot topic of research but most of the researches/findings have not been used in the production environments.

Intrusion detection system is a mechanism that helps in detecting unwanted intrusions. Intrusion is when someone unauthorised or something unwanted enters in our system. The concept of intrusion detection was coined by Anderson in 1980 (Anderson,

1980) and formalised by Denning in 1987 (Denning, 1987). IDS detects such intrusions using various techniques. An IDS can be a host-based IDS or a network-based IDS (Sarmah, 2001). Host-based intrusion detection system is installed on host to monitor the traffic coming to and from the host whereas network-based intrusion detection system is installed on certain points in a network and monitor the traffic data passing through those points. Network intrusion detection systems are more capable of detecting wide range of intrusions. Network-based IDS can further be categorised into misuse-based and anomaly-based. In misuse-based IDS, signatures of unwanted intrusions/misuses are stored in a database. To term an activity on a network an intrusion, it must match with an entry in the misuse database. In an anomaly-based intrusion detection, accepted behaviour of a system is modelled and any activity on the network which falls outside the predefined or accepted behaviour is termed as anomalous.

1.1 Anomaly detection

In Denning (1987), Denning devised a real-time intrusion detection expert system in which he hypothesised that security violations could be detected from abnormal pattern of system usage, i.e., anomalies in the behaviour of a system can be detected. Anomaly-based IDS has certain advantages over misuse-based IDS which are listed below:

- In the misuse-based IDS only known intrusions, which are there in signatures database, can be detected. Any new intrusion or zero day intrusions cannot be detected. As anomaly-based IDS does not require prior knowledge of intrusions, new anomalies or zero day anomalies can be detected more easily.
- Scalability is not an issue in anomaly-based IDS as zero day intrusions are detectable. However, in misuse-based IDS, new intrusions can be detected only if an entry is made for that intrusion in the misuse database.

Though signature-based intrusion detection systems are widely used as compared to anomaly-based intrusion detection systems because of the ease of use and availability, yet anomaly-based intrusion detection systems are gaining momentum as these are more promising than misuse-based intrusion detection systems. The most important step in anomaly-based intrusion detection is modelling the acceptable behaviour of the system. Accuracy of results obtained largely depends on the behavioural model of the system. Various intrusion detection systems are used now-a-days which include open source IDS like Snort (Beale et al., 2007), Suricata (Day and Burns, 2011), IDS Bro (Paxson, 1999) (http://www.bro-ids.org), OpenWIPS-ng (http://www.openwips-ng.org) and Sguil (Bianco, 2005); commercial IDS like metaflows security systems from Metaflows Inc. (http://www.metaflows.com), SecureWorks managed IDS/IPS (SecureWorks, 2014) (https://www.secureworks.com/resources/sb-advanced-threat-protection), Symantec managed IDS/IPS from Symantec (2005), CounterSnipe IPS from CounterSnipe (http://www.countersnipe.com), Dragon by Enterasys (Allan, 2002) (http://www.bus.umich.edu/kresgepublic/journals/gartner/research/105000/105094/105094.html), FireStorm NIDS (Leach and Tedesco, 2003) (http://www.scaramanga.co.uk/firestorm/documentation.html), etc.

1.2 IP flows

An intrusion detection system can work on one of the following three types of data available in a computer network:

- packet data
- flow records
- traffic aggregation at autonomous system (AS).

Packet data is the detailed data used for intrusion detection which is also termed as deep packet inspection (DPI) but many a times DPI is not possible due to various reasons like privacy, encryption and complexity. As the network traffic is increasing in volume and complexity, it is quite hard to capture and decode every packet for analysis. Due to these limitations research fraternity is moving towards Flow-based intrusion detection in which flow records are inspected to find intrusions. A flow is a combination of number of packets sharing source IP, destination IP, source port, destination port and protocol. Only header data is used to create a flow, therefore, privacy and encryption do not affect the outcomes. Moreover, different levels of analysis are possible and as the flow-based data is small in size as compared to packet level data for the same traffic, it is very suitable for high speed networks. Traffic can be aggregated further at AS level but it hides lots of data which could otherwise be useful.

There are certain papers that give overview of network monitoring for intrusion detection and performance monitoring from different aspects (Hofstede et al., 2014; Marnerides et al., 2014; Sperotto et al., 2010; Li et al., 2013). No paper discusses specific techniques and methods used by researchers for handling the problem of flow-based anomaly detection in computer networks. The aim of this paper is to present an overview of the methods and techniques employed by researchers for detecting anomalies in the system based on network flows. This survey is based on review of latest state of the art techniques/algorithms developed by researchers in recent times. Another important aim of this paper is to identify open problems and research challenges in the field of flow-based anomaly detection. The remainder of this paper is organised as follows. Section 2 gives overview of various flow-based datasets available for performing experiments to validate the proposed techniques and methods. Section 3 gives detailed overview of techniques and methods used by researchers for anomaly detection in networks. Open problems and research challenges are discussed in Section 4 followed by conclusion in Section 5.

2 Flow-based datasets

Not many flow-based datasets are available and those which are available are having certain limitations. We are giving an overview of three flow-based datasets which are containing data records in the form of flows only.

2.1 Flow-based Tezpur University dataset for intrusion detection system

Gogoi et al. (2012) proposed new real packet level dataset and flow level dataset. A test bed is setup for collecting data which included router, L3 and L2 switches, server, workstations and nodes. Server had a mirror port to observe traffic. Packet level traffic is captured using gulp and analysed using Wireshark. Fifty features are extracted using tcptrace which are classified into basic, content-based, time-based and connection-based. Netflow standard is used for flow collection. Nf dump is used to collect flows with the help of daemon process nfcapd. Files are saved at 5 minutes interval and C programs are used to filter data and extract 24 new features which are classified into basic, time window-based and connection-based classes. Sixteen different types of attacks are generated using attack generation tools like targa, nmap, rnmap, brute force ssh, agent handler network and IRC botnet which are incorporated into the dataset. Datasets have been validated using number of existing techniques. This dataset is no more available online or otherwise.

2.2 Flow-based dataset by Sperotto

Sperotto et al. (2009) proposed first flow-based labelled dataset for intrusion detection. The data is collected at University of Twente, The Netherlands. A honeypot was installed in a virtual machine for six days and the host was configured by installing ssh, Apache webserver and ftp services on the host. Log file was created and this log file was monitored to identify attacks and label the flows. Around 14 million flows were collected. Flow creation is done using softflowd. Cluster alerts are also created to label logical groups of alerts. This dataset contains malicious records only, normal data is missing in this dataset and more than 98% flows are labelled.

2.3 Winter dataset

Winter et al. (2011) modified flowbased dataset by Sperotto (Sperotto et al., 2009) by discarding IP addresses as they were anonymised, unlabeled flows were deleted and all the flows belonging to protocols other than SSH and HTTP were deleted. Random sampling was performed and at the end 22,924 flows were left which were used in the study by authors. As Sperotto dataset did not contain benign data, benign data was generated belonging to HTTP, SSH, DNS, ICMP and FTP. Benign data comprises of 1,904 flows.

3 Flow-based anomaly detection

This section gives an overview of flow-based anomaly detection techniques used by researchers for tackling the problem of anomaly-based intrusion detection. Flow-based anomaly detection depends on the modelling of normal behaviour of flows and detecting deviations from the normal behaviour. Flow-based monitoring is gaining momentum because of the advantages it is offering. Steps of general architecture of flow monitoring are given below (Hofstede et al., 2014).

- packet observation
- flow metering and export
- data collection
- data analysis.

Various open source flow exporters like ipt-netflow, softflowd, nProbe, pmacct, QoF, Vermont, YAF and commercial flow exporters like FlowMon Probe, Stealth Watch Flow Sensor, nBox, Junos JFlow, etc. are available many of which supports Netflowv5, Netflowv9 and IPFIX. Similarly many open source FlowCollectors like Argus, flowd, nfdump, nProbe, pmacct, SiLK, vermont and commercial flow collectors like Arbor networks, Fluke networks, INVEA-TECH, Plixer, solar winds are available for use. Various commercial data analysis applications like Fluke networks, Compuware, Lancope, Plixer, SevOne and open source data analysis applications like FlowViewer, NfSen, Stager, nTop, etc. are available for analysing the collected data. Detailed comparison is available in Hofstede et al. (2014). Flow-based detection is the need of the hour as only header information is taken into consideration instead of packet contents that helps in taking action in real-time. Flow-based detection cannot handle some classes of attacks like semantic-based attacks. So flow-based detection is complimentary and not replacement for DPI methods (Sperotto et al., 2010). Moreover packet sampling and flow sampling directly affect processing costs and traffic features (Marnerides et al., 2014). Flowbased data analysis started in last years of 20th century and the interest is increasing day by day (Li et al., 2013).

The techniques and methodologies applied by researchers on flow-based network traffic data based on the papers covered in this overview are categorised in five classes:

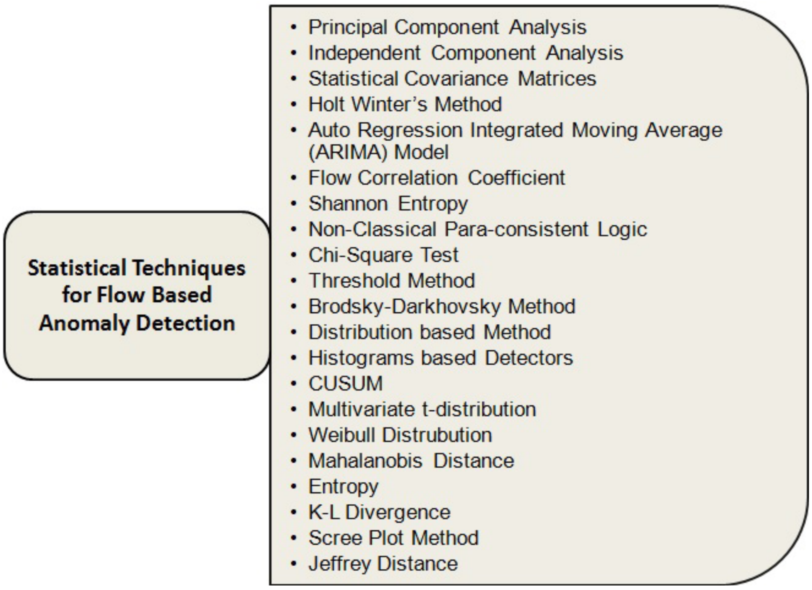
- statistical techniques
- machine learning techniques
  - a neural network-based techniques
  - b support vector machines-based techniques
  - c meta-heuristic techniques
- clustering techniques
- frequent pattern mining techniques
- agent-based techniques.

3.1 Statistical techniques for anomaly detection

Statistics is the study of the collection, analysis using various standard techniques, interpretation, presentation, and organisation of data. A large number of statistical techniques are used for flow-based anomaly detection as shown in Figure 4.

The most basic and fundamental method of finding anomalies in network is to monitor quantative attributes like volume of flows against thresholds. Simply comparing values with thresholds does not give convincing results; therefore, people are employing specific statistical methods for generating thresholds and comparing values. Ellens et al. (2013) used flow-based variables to identify anomalies by using three methods: threshold method, Brodsky-Darkhovsky method and distribution-based methods for monitoring peaks in traffic time series, changes in average amount of traffic and changes in the underlying traffic distribution. Detection rate and number of false positives per session metrics are used to evaluate the performance which confirms that diversified tunnel usage scenarios can be identified with high detection rate. Karasaridis et al. (2006) used a threshold value on the number of flows to generate alerts. Alerts for repeated requests, repeated responses and cache poisoning attacks are generated by matching the request response flows. Alerts for tunnelling attacks are generated by monitoring packet size statistics. Frequency of packets not conforming to normal protocol size is used for requests as well as responses for detecting the attacks. Cross entropy-based anomaly detection is used over packet size, histograms and CUSUM-based detectors are used over frequencies of non-conforming packets. In cross entropy detection, observed probabilities are compared to baseline probabilities to detect anomalies. Roy et al. (2017) proposed a distributed approach for the detection of dumb nodes in wireless sensor network which can also be considered as an anomalous condition. In the proposed approach, cumulative sum (CUSUM) approach is used to detect anomalous condition and the detection problem is analysed using Markov chain.

Figure 4 Statistical techniques for flow-based anomaly detection


Statistical Techniques for Flow Based Anomaly Detection
• Principal Component Analysis
• Independent Component Analysis
• Statistical Covariance Matrices
• Holt Winter's Method
• Auto Regression Integrated Moving Average (ARIMA) Model
• Flow Correlation Coefficient
• Shannon Entropy
• Non-Classical Para-consistent Logic
• Chi-Square Test
• Threshold Method
• Brodsky-Darkhovsky Method
• Distribution based Method
• Histograms based Detectors
• CUSUM
• Multivariate t-distribution
• Weibull Distrubution
• Mahalanobis Distance
• Entropy
• K-L Divergence
• Scree Plot Method
• Jeffrey Distance

Principal component analysis (PCA) is a widely used statistical technique for flow-based anomaly detection. It is a statistical procedure that uses orthogonal transformations to convert possibly correlated variables into linearly uncorrelated variables called principal components. Fernandes et al. (2015) uses PCA technique for profiling normal behaviour of network using quantitative flow-based attributes. Again in Fernandes et al. (2016), Fernandes used PCA along with ant colony optimisation metaheuristic to generate traffic profile for normal network behaviour which can be compared with real network traffic for detecting anomalous events. PCA technique has also been used with Haar wavelet analysis in the Hybrid approach proposed in Novakov et al. (2012). In the hybrid approach, modified PCA, that includes time shift feature, generates principal components and the most significant component is used by Haar wavelet analysis to measure degree of change called deviation score. Deviation score is calculated for individual time bins and absolute difference in deviation scores of consecutive bins is used to find out potential anomaly. PCA is also used by Callegari et al. (2011) to detect anomalies and identify the flows at the aggregated level which are responsible for anomaly. In the method, netflow data is collected over time bins which is converted into matrix form by first applying entropy and then K-L divergence. Dominant and negligible principal components are generated and partitioned into normal and anomalous subspaces. Sketches are used to map data into smaller set and scree-plot method is used to find out optimal number of principal components.

Normal TCP session arrivals are well described using Poisson processes (Floyd and Paxson, 2001). Arshadi and Jahangir (2011) shows that inter arrival times of TCP flows conform to Weibull distribution and not Poisson distribution. By estimating Weibull parameters, normal traffic can be modelled and any deviation from Weibull distribution confirms the presence of anomalous flows. Weibull parameters are estimated using median rank method and Chi-square test is applied as a measure of goodness of fit to confirm the data against Weibull distribution. For anomaly detection SYN packets are divided into non-overlapping consecutive windows and Weibull parameters of inter arrival times are estimated in each window. Conformity of data is checked using Chi-square test. Packets in the windows with high level of significance show anomalous behavior which can be analysed further. Li et al. (2015) considered the problem of anomaly detection in the presence of noise interference and data loss. They used multivariate t-distribution of data for normal traffic modelling and latent variable probability theory is applied in order to simplify parameter estimation and handling missing data. Origin destination flows are maintained as traffic matrix using sample's Mahalanobis distance that exceeds threshold, then contribution analysis is performed for anomaly localisation.

Assis et al. (2013) focuses on qualitative attributes rather than quantitative attributes. The qualitative attributes are converted into quantitative values using Shannon entropy and then Holt Winters method based on exponential weighted moving average (EWMA) method is used to characterise baseline, linear trend and seasonality. An equation is generated based on these three characteristics is used to forecast the values for next intervals. Confidence bands are also generated and updated dynamically using previous interval values. The confidence bands give interval in which deviations are considered normal. Another approach based on moving averages is used in Pena et al. (2014) where authors presented an approach in which digital signature of network segment is generated using autoregressive integrated moving average (ARIMA) model as the time series data under study, are non-stationary. Paraconsistent annotated logic with annotation of two values (PAL 2v), an extension of non-classical paraconsistent logic is used to take care of the inconsistent information which leads to uncertainty. Degrees of faith and discredit are used as evidence for anomaly. Concept of confidence bands are also used, Yan (2016) analysed differences between Renyi entropy and Shannon entropy and found Renyi entropy better than Shannon entropy in terms of overall accuracy, average precision and average true positive rate. EWMA control chart theory is used to detect and screen anomalies with the help of features measured using Renyi entropy. A lightweight information-entropy based metric is proposed by Bhuyan et al. (2016b) for detecting DDoS flooding attacks. The proposed system uses generalised entropy metric with packet intensity computation to measure the metric difference between legitimate traffic and attack traffic. A lightweight IP traceback scheme-based on entropy metric is also proposed that can trace the location of attacker's machine.

Yu et al. (2012) proposed a discrimination algorithm for detecting DDoS-based anomalies specifically using the fact that DDoS attack flows are more similar to each other than to flash crowd flows. For suspected pair of flows, a flow correlation coefficient is defined which can be compared against a threshold for discrimination to indicate the presence and absence of attack. Phase difference between correlated flows can indicate the absence of correlation. This situation has been handled by introducing position shifts in flows before computing correlation coefficient. Normalised mean square error, correlation coefficient and symmetric mean absolute percentage error are used by authors to test the effectiveness of profile generated. Receiver operation characteristics (ROC) graphs based on true positive rate and false positive rate, accuracy and precision are used by authors to measure effectiveness of detection systems. Li et al. (2014) focused on defense scheme that adapts and responds autonomously to DDoS attacks. A distributed defense scheme based on two stage traffic flow control is proposed which deploys coordinated modules to protect servers.

Traffic data involved in anomaly detection is quite large in amount for which, sketch data structures are quite useful. Sketch data structure is an array of hash tables which contains counters used to randomly aggregate flow records. Salem et al. (2012) used data structure in which counters are used to establish a probabilistic model for traffic. Sliding Window model is used to handle large amount of data in high speed networks. A dynamic threshold value is used to compare the divergence and report anomalies. Johnson (Johnson and Lazos, 2014) investigated the use of traffic aggregation at AS level for anomaly detection and found that the aggregation mitigates storage and computational scalability problems. The NIDS operates at border gateway servicing traffic and works in three phases: data aggregation, statistical analysis and anomaly detection. Empirical probability distribution using count-based histograms is created during training phase and comparison of online distribution is done with distribution of training phase to find distance using Jeffrey distance.

Other Statistical techniques used by researchers include independent component analysis (ICA) and matrices-based method. Palmieri et al. (2014) used ICA for generating normal profile and supervised decision tree-based learning algorithm C4.5 to train the system. Statistical covariance matrices are used to build normal profiles by Yeung et al. (2007). Each element in the covariance matrix represents covariance between two features of time series traffic data. Chebyshev inequality theorem is used to determine threshold matrix which helps in detecting significant differences between normal profiles and deviating profiles.

3.2 Machine learning-based techniques for anomaly detection

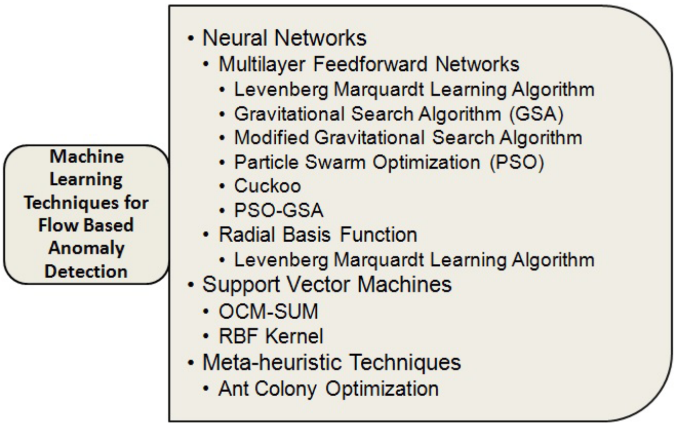
Machine learning techniques are the emerging techniques used for flow-based anomaly detection. Various machine learning techniques like neural networks, support vector machines, meta-heuristic techniques, classification techniques are becoming the focus of attention for anomaly detection as shown in Figure 5.

3.2.1 Neural network-based techniques

Artificial neural networks are a family of models inspired by biological neural networks which are used to approximate functions which are generally unknown. Neural networks-based techniques are prominently used by authors for anomaly detection.

An intrusion detection system based on two neural network stages is proposed by Abuadlla et al. (2014) where in the first stage significant changes in the network are detected and in the second stage, known attack types are used to classify detected anomalies. Multilayer feedforward neural network and radial basis function network are used with Levenberg-Marquorot training algorithm. The system was designed to detect specific attack types Dos/DDoS, port scan, land attack and other attacks. Radial basis function network is concluded better for real-time detections. Another system using multilayer perceptron is proposed by Sheikhan and Jadidi (2014) in which modified gravitational search algorithm and particle swarm optimisation (PSO) algorithms are used to train the system. Modified gravitational search algorithm gives better results. Jadidi et al. (2013a) used a multilayer perceptron with a single layer of hidden neurons. Two meta-heuristic algorithms: Cuckoo and PSOGSA which is a combination of PSO and GSA algorithms are used. System is designed in MATLAB. PSOGSA algorithm is concluded better than Cuckoo algorithm. Jadidi et al. (2013b) designed a system using two layer multilayer perceptron which is trained using gravitational search algorithm as it has faster convergence and adaptive learning rate and it is found better than PSO algorithm. Metrics used by authors for measuring the performance are correct classification rate, error rate, miss rate, false rate and accuracy.

Figure 5 Machine learning techniques for flow-based anomaly detection


Machine Learning Techniques for Flow Based Anomaly Detection
• Neural Networks
• Multilayer Feedforward Networks
• Levenberg Marquardt Learning Algorithm
• Gravitational Search Algorithm (GSA)
• Modified Gravitational Search Algorithm
• Particle Swarm Optimization (PSO)
• Cuckoo
• PSO-GSA
• Radial Basis Function
• Levenberg Marquardt Learning Algorithm
• Support Vector Machines
• OCM-SUM
• RBF Kernel
• Meta-heuristic Techniques
• Ant Colony Optimization

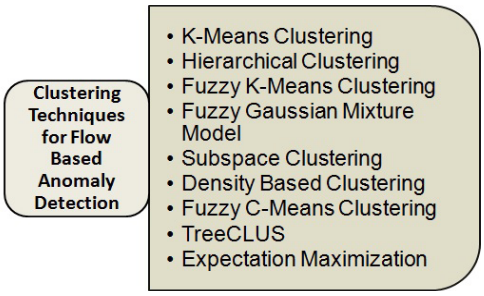
3.2.2 Support vector machines-based techniques

Support vector machines are a set of related supervised learning methods that analyse data and recognise patterns. These machines are used for classification and regression analysis. Winter et al. (2011) used an unsupervised one class support vector machine for designing a classification system which is trained for malicious data and not the benign one. Radial basis function kernel was used because of its general applicability and 98% correct prediction is achieved. Ippoliti and Zhou (2014) presented an integrated approach using adaptive support vector machine in which tradition flows are augmented with additional features computed in an online fashion using count-min sketches. These augmented flows are used to directly extract information related to volume and distribution. One class SVMs are used with several enhancements which are used in an integrated fashion. 95% precision and recall are achieved in the proposed approach.

3.2.3 Metaheuristic techniques

For normal profile generation, Carvalho et al. (2013) used a modification of Ant colony optimisation meta-heuristic technique. Adaptive timewarping pattern matching technique is used to compare real-time behaviour and generated profiles. It works in two stages, in the first stage, conventional dynamic time warping is used to find similarity calculation which considers shape and in second stage, distance is calculated considering amplitude. ROC curve is used to measure efficiency of the system. Sheikhan and Jadidi (2014) used modified gravitational search algorithm and PSO meta-heuristic algorithms to train multilayer perceptron for detecting anomalies. Modified gravitational search algorithm gives better results.

Figure 6 Clustering techniques for flow-based anomaly detection


Clustering Techniques for Flow Based Anomaly Detection
• K-Means Clustering
• Hierarchical Clustering
• Fuzzy K-Means Clustering
• Fuzzy Gaussian Mixture Model
• Subspace Clustering
• Density Based Clustering
• Fuzzy C-Means Clustering
• TreeCLUS
• Expectation Maximization

3.3 Clustering techniques for anomaly detection

Clustering is a technique of grouping set of objects in such a way that objects in the same group, which is called a cluster, are more similar to each other than to those in other groups/clusters. The clustering techniques used by researchers for network anomaly detection are shown in Figure 6.

Zang et al. (2011) studied the use of hierarchical and K means clustering for detecting botnets specifically. Botnet traffic generated in the experimental setup is mixed with normal traffic and then filtered by keeping TCP-based flows only. Fifteen derived features are used. Experimental results showed that k-means clustering can reach perfect performance and consumes much less time.

Bhuyan et al. (2012) presented an outlier-based approach for coordinated port scan detection. PCA technique is used for feature extraction followed by fuzzy cmeans clustering to cluster each sample into k clusters. Range-based profile is generated for each cluster, then profiles are matched with others to remove redundancy. Outlier score is calculated and those having value higher than user defined threshold are termed anomalous. Adaptive updation of clusters is done to increase effectiveness. Bhuyan et al. (2016a) presented multi-step outlier-based approach in which a subset of traffic features are identified using mutual information and generalised entropy-based feature selection technique. Tree-based clustering technique is used to generate reference points which are used by an outlier score function that ranks incoming network traffic to identify anomalies.

Casas et al. (2011) introduced an unsupervised network anomaly detection algorithm (UNADA) in which flow aggregation is done on different keys. Change detection is performed and then clustering is done over source IPs and destination IPs. An ensemble clustering approach is used which combines subspace clustering and density-based clustering DBSCAN algorithm. Clustering results are passed on to Evidence Accumulation for ranking outliers which constructs dissimilarity vector that stores distance between different outliers, weighing factor w is used for boosting the outlier parameter and Mahalanobis is used as a measure of dissimilarity.

A hierarchical intrusion detection system based on a binary tree is proposed by Ahmim and Zine (2015). At each level of binary tree, a classifier that classifies network connections into two classes is used, where one class represents the particular type of connections and other class represents all other connections. At next level, again a classifier is applied on rest of the connections of previous level to classify them further into a known connection class and a group of rest of the connections. Selection of classifier at each level is done on the basis of study performed on the dataset. Principle followed for classification in the proposed IDS is highest performance with different categories of attacks. For comparing classifiers performance, detection rate, false alarm rate, global detection rate and accuracy metrics are used.

A framework, STONE is proposed by Gulisano et al. (2015) that creates normal profiles of services by aggregating SourceIPs into common prefixes and creating clusters of flow records on the basis of aggregated source IPs. This aggregation is done as the aggregated behaviour is more stable than individual behaviour. These clusters are grouped into various groups depending on traffic features and ratio of clusters for each group is maintained. Whenever an abrupt change is observed in this ratio, an anomaly is detected. After detecting anomaly, mitigation is done by filtering out anomalous traffic data. STONE is implemented on top of Stream Cloud which is a distributed stream processing engine that works on sliding window principle of handling stream of data.

Liu et al. (2013) used fuzzy k-means and Gaussian mixture model (GMM) for classification. Non-negative matrix factorisation (NMF) and PCA are used for feature transformation. It is concluded that feature selection and reduction are more appropriate than feature transformation and PCA can extract more effective features but computation overhead is more and fuzzy GMM is more robust than k-means.

Bhuyan et al. (2011) presented a clustering and outlier-based approach for network anomaly detection which consists of four steps: feature selection, clustering, profile generation and outlier detection. Information gain-based feature selection is used for identifying minimum set of features. Clustering technique TreeCLUS is proposed and profile is generated for each cluster. Larger clusters are considered normal and smaller clusters are considered outliers.

Liu et al. (2016) introduced PSOLGA which is a PSO-based clustering algorithm to create behaviour profiles for each server application by mining significant linear structures in flows. PSOLGA is a combination of PSO and linear grouping algorithm (LGA) which combines ideas from principal components, clustering methods and resampling algorithms.

Fachkha et al. (2015) presented an approach based on darknet space. A flow of traffic is considered anomalous because of DNS amplification DDoS attack if atleast 21 DNS query packets of type ANY are sent to atleast 29 distinct unused dark IP addresses. Expectation maximisation (EM) and K-means clustering are used to cluster similar DNS amplification DDoS traces in order to find out campaigns of attacks.

3.4 Frequent pattern mining-based techniques for anomaly detection

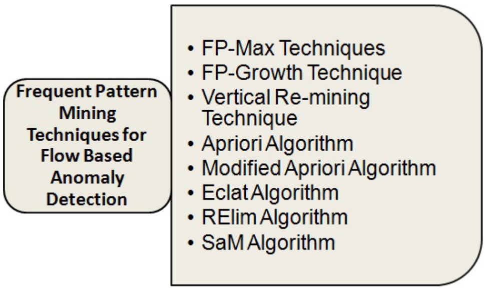
Frequent Pattern is a pattern that occurs frequently in the data. Mining a frequent pattern means to find significant frequent pattern which could be useful for drawing important conclusions. Researchers are exploring the use of frequent pattern mining/association rule mining for anomaly detection and finding good results. The techniques used are shown in Figure 7.

Paredes-Oliva et al. (2012) proposed a system to detect and classify classes of anomalies. Frequent combinations of features are extracted using FP max frequent pattern mining technique, which are given to a C5.0 classification tree for automatic detection.

Li and Deng (2010) proposed four frequent pattern mining algorithms: vertical re-mining, multi-pattern re-mining, fast multi-pattern capturing and fast multipattern capturing supplement for finding frequent patterns. These algorithms use vertical representation of database to generate patterns. SlidingWindow concept is used to process stream data where a sliding window is composed of multiple basic windows such that as frequent patterns for next basic window are derived, patterns of outdated basic window are discarded. A tree structure is used to maintain frequent patterns generated for basic windows.

Brauckhoff et al. (2012) used n different histogram-based anomaly detectors in an online fashion to detect anomaly and trigger the alarm. Metadata generated by these histogram-based techniques are combined using union operator. Suspicious flows are made to pass through the union of metadata such that normal flows are filtered out. Frequent patterns are generated using modified Apriori technique for the anomolous flows which can be analysed manually to find root cause of detected anomalies.

Figure 7 Frequent pattern mining techniques for flow-based anomaly detection



Frequent Pattern Mining Techniques for Flow Based Anomaly Detection

- FP-Max Techniques
- FP-Growth Technique
- Vertical Re-mining Technique
- Apriori Algorithm
- Modified Apriori Algorithm
- Eclat Algorithm
- RElim Algorithm
- SaM Algorithm

Paredes-Oliva et al. (2013) developed fast recognition of high multi-dimensional network (FaRNet) for network traffic profiling. The system discovers hierarchical multi-dimensional network traffic heavy hitters using the technique of frequent item set mining. Five FIM algorithms Apriori, FP-Growth, Eclat, RElim and SaM are used and evaluated. Effect of sampling is also studied and it is concluded that sampling will not have negative impact on results if sampling rate is decided wisely which can be derived using binomial expansion. The hierarchical nature of various attributes like IP addresses are exploited using full expansion and progressive expansion k by k.

3.5 Agent-based techniques for anomaly detection

An Intelligent Agent is a piece of software that functions as an agent for user or another software. Agents can activate and run themselves. Agent-based approach has been proposed by Rehak et al. (2008) who developed an agent-based system in which each agent uses a specific anomaly detection method to detect anomalies. Each flow is assigned an anomaly value depending upon the level of anomaly detected in it. These anomaly values for individual flows are integrated into their trust models by the agents. An aggregation agent takes the trustfulness of each flow from different agents for estimating its maliciousness. This estimation is used for traffic filtering and visualisation.

Rehak et al. (2009) improved its system (Rehak et al., 2008) by making it adaptive. Along with detection agents and aggregation agents, incident agents are also used which can migrate between containers containing detection and aggregation agents. Incident agents represent knowledge about specific incidents including known attacks or legitimate traffic types which is used in anomaly detection stage for computing final trust estimates by the aggregation agents which are ranked and best aggregation agent is selected. Similarly threshold values separating malicious and legitimate traffic are also determined and each aggregation agent receives feedback from incident agent for adaptive updating.

Roy et al. (2014) proposed a mobile agent-based approach to detect dumb nodes in wireless sensor networks which can also be considered as an anomalous condition. Dempster Shafer theory is used to assign reward and penalty frequencies to the current node by the neighbouring nodes which helps in predicting the actual behaviour of each node and hence detecting dumb nodes.

Boukhlouf et al. (2016) proposed a distributed intrusion detection system in which platform aglets are used for creating and distributing four types of mobile agents which detect intrusions in the system.

Kumar and Venugopalan (2017) establishes the fact that if the dataset is splitted initially-based on ‘protocol type’ feature, before applying any technique for anomaly detection, the performance w.r.t. rate of detection and time to build model is enhanced. Using more than one technique iteratively can also help in increasing accuracy. Ahmim and Ghoualmi-Zine (2014) used two different classifiers iteratively to detect low frequent attacks along with high frequent attacks.

4 Research problems and open challenges

Computer networks are facing new attacks and problems on daily basis. Existing signature-based systems can detect those anomalies which are well known and registered in signature database. Zero day attacks are difficult to detect and mitigate using these systems. Anomaly-based systems are good answer to the problems associated with misuse-based systems but anomaly-based systems are not 100% accurate as they model normal behaviour and detect deviations. Flow-based systems are based on header information of the packets, so they also do not provide 100% accuracy. The accuracy/detection rate achieved by some of the state-of-the-art techniques proposed in literature is shown in Table 1. However flow-based systems are need of the hour because of the increasing speed and size of networks day by day. In this section, we will discuss open problems and research challenges associated with flow-based anomaly detection Systems.

- Flow-based systems work on flow-based traffic data collected from a network. Flow-based data includes data available in headers of packets. Netflow v5, Netflow v9 and IPFIX are standards for flow-based data. Netflow v5 supports 5-tuple records containing SourceIP/DestinationIP, SourcePort/DestinationPort and protocol. Netflowv9 and IPFIX supports extended data to be included which includes application awareness also. Explicit flowbased datasets are not available for verifying the systems. Three flowbased datasets are identified and discussed in Section 2.2 but each dataset has some problem associated with it. So a duly labelled benchmark flowbased dataset is the need of the hour which should represent real scenarios.
- Existing anomaly-based systems are not 100% accurate. False positives and false negatives are high in number in existing systems. It is a real challenge to increase the accuracy of the system by decreasing false positives and false negatives. New techniques and methods need to be identified which can serve the purpose.
- Networks are controlling virtually everything in this world, so a network needs to be working always. Any problem encountered in the network should be detected and tackled in real-time to avoid the total shut down of the network. Real-time detection and mitigation in flowbased anomaly detection is another open challenge for researchers.
- For mitigating the detected problem, exact root cause of the problem is to be identified. In many cases, root cause is missed out or it is to be found by manually inspecting the portion of the network in which anomaly is detected. Automating the detection of root cause of Anomaly is another important research problem for flow-based anomaly detection.
- New problems are emerging day by day in networks. Many such problems like DNS tunnelling, SQL injection, etc. goes undetected. Understanding the anomalous behaviour produced by these problems is very important in order to incorporate them into the detection system.

Table 1 Comparison of state-of-the-art techniques

Research paper	*Class of technique(s) used*	*Anomalies/attacks*	*Accuracy/detection rate*
Yeung et al. (2017)	Statistical	DDoS flooding attacks	88.89% (for normal) 100% (for attacks)
Callegari et al. (2011)	Statistical	DoS and other anomalies	85%
Winter et al. (2011)	Machine learning	SSH and HTTP anomalies	98.07%

Table 1 Comparison of state-of-the-art techniques (continued)

Research paper	*Class of technique(s) used*	*Anomalies/attacks*	*Accuracy/detection rate*
Bhuyan et al. (2011)	Clustering	R2L	87.76%
		DoS	99.98%
		Probe	95.70%
		U2R	68.42%
Paredes-Oliva et al. (2012)	Frequent pattern mining	DoS, port scan network scan unknown	98%
Bhuyan et al. (2016a)	Clustering	R2L	89.96%
		DoS	99.99%
		Probe	98.07%
		U2R	76.32%
Bhuyan et al. (2012)	Clustering	Probe	99.02%
Jadidi et al. (2013a)	Machine learning	Anomalies	99.55%
Jadidi et al. (2013b)	Machine learning	Anomalies	99.43%
Carvalho et al. (2013)	Machine learning	DDoS	92%
Assis et al. (2013)	Statistical	Anomalies	85%
Pena et al. (2014)	Statistical	Anomalies	90.03%
Sheikhan and Jadidi (2014)	Machine learning	SSH scanning, DoS	97.8%
Palmieri et al. (2014)	Statistical	SYN floods	99.38%
Ippoliti and Zhou (2014)	Machine learning	Anomalies	95% (precision) 95% (recall)
Abuadlla et al. (2014)	Machine learning	DoS attack	100%
		Port scan attack	99.9%
		Land attack	100%
		Unknown attack	78%
Fachkha et al. (2015)	Clustering	DRDoS	82%
Fernandes et al. (2015)	Statistical	DoS, DDoS and flash crowds	85%
Ahmim and Zine (2015)	Classification	DoS attack	99.45%
		Probe attack	84.11%
		R2L attack	36.17%
		U2R attack	8.77%
Liu et al. (2016)	Machine Learning	Sql injection attempts	98.45%
Fernandes et al. (2016)	Statistical and machine learning	DoS, DDoS, flash crowds	96%
Bhuyan et al. (2016b)	Statistical	DDos flooding attacks	99.77%

5 Conclusions

This review paper has discussed flow-based anomaly detection in depth, justifying why flow-based anomaly detection is the need of the hour. Various flow-based datasets have been reviewed in the paper. The techniques used by researchers to tackle the problem are classified and reviewed in detail. Flow-based anomaly detection has not been accepted fully to ensure the security of the system. Various open challenges and research problems associated with flow-based anomaly detection have also been discussed in detail.

References

Abuadlla, Y., Kvascev, G., Gajin, S. and Jovanovic, Z. (2014) 'Flow-based anomaly intrusion detection system using two neural network stages', *Computer Science and Information Systems*, Vol. 11, No. 2, pp.601–622.

Ahmim, A. and Ghoualmi-Zine, N. (2014) 'A new adaptive intrusion detection system based on the intersection of two different classifiers', *International Journal of Security and Networks*, Vol. 9, No. 3, pp.125–132.

Ahmim, A. and Zine, N.G. (2015) 'A new hierarchical intrusion detection system based on a binary tree of classifiers', *International Journal of Information & Computer Security*, Vol. 23, No. 1, pp.31–57.

Allan, A. (2002) *Enterasys Networks Dragon Intrusion Detection System* [online] http://www.bus.umich.edu/kresgepublic/journals/gartner/research/105000/105094/105094.html (accessed 10 May 2016).

Anderson, J.P. (1980) *Computer Security Threat Monitoring and Surveillance*, Technical report, James P. Anderson Company, Fort Washington, Pennsylvania.

Arshadi, L. and Jahangir, A.H. (2011) 'On the TCP flow inter-arrival times distribution', *Fifth UKSim European Symposium on Computer Modeling and Simulation (EMS)*, pp.360–365, IEEE.

Assis, M., Rodrigues, J. and Proenca, M.L. (2013) 'A hybrid approach for anomaly detection on large-scale networks using HWDS and entropy', *21st International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, pp.1–5, IEEE.

Beale, J., Baker, A.R. and Esler, J. (2007) *Snort: IDS and IPS Toolkit*, Syngress, Burlington, MA.

Bhuyan, M.H., Bhattacharyya, D.K. and Kalita, J.K. (2011) 'NADO: network anomaly detection using outlier approach', *Proceedings of the 2011 International Conference on Communication, Computing & Security*, pp.531–536, ACM.

Bhuyan, M.H., Bhattacharyya, D.K. and Kalita, J.K. (2012) 'AOCD: an adaptive outlier based coordinated scan detection approach', *International Journal of Network Security*, Vol. 14, No. 6, pp.339–351.

Bhuyan, M.H., Bhattacharyya, D.K. and Kalita, J.K. (2016a) 'A multi-step outlier-based anomaly detection approach to network-wide traffic', *Information Sciences*, Vol. 348, No. C, pp.243–271.

Bhuyan, M.H., Bhattacharyya, D.K. and Kalita, J.K. (2016b) 'E-LDAT: a lightweight system for DDoS flooding attack detection and IP traceback using extended entropy metric', *Security and Communication Networks*, Vol. 9, No. 16, pp.3251–3270.

Bianco, D.J. (2005) *Open Source Network Security Monitoring with Sguil* [online] http://www.vorant.com/files/nsm_with_sguil.pdf (accessed 8 May 2016).

Boukhlouf, D., Kazar, O. and Kahloul, L. (2016) 'Network security: distributed intrusion detection system using mobile agent technology', *International Journal of Communication Networks and Distributed Systems*, Vol. 17, No. 4, pp.335–347.

Brauckhoff, D., Dimitropoulos, X., Wagner, A. and Salamatian, K. (2012) 'Anomaly extraction in backbone networks using association rules', *IEEE/ACM Transactions on Networking (TON)*, Vol. 20, No. 6, pp.1788–1799.

Callegari, C., Gazzarrini, L., Giordano, S., Pagano, M. and Pepe, T. (2011) 'A novel PCA based network anomaly detection', *International Conference on Communications (ICC)*, pp.1–5, IEEE.

Carvalho, L.F., Rodrigues, J., Barbon, S. and Lemes, M. (2013) 'Using ant colony optimization metaheuristic and dynamic time warping for anomaly detection', *SoftCOM*, pp.1–5.

Casas, P., Mazel, J. and Owezarski, P. (2011) 'Unada: Unsupervised network anomaly detection using sub-space outliers ranking', *International Conference on Research in Networking*, pp.40–51, Springer Berlin Heidelberg.

CounterSnipe IPS [online] http://www.countersnipe.com (accessed 10 May 2016).

Day, D. and Burns, B. (2011) 'A performance analysis of snort and suricata network intrusion detection and prevention engines', *Fifth International Conference on Digital Society*, Gosier, Guadeloupe, pp.187–192.

Denning, D. (1987) 'An intrusion-detection model', *IEEE Transactions on Software Engineering*, Vol. SE-13, No. 2, pp.222–232.

Ellens, W., Zuraniewski, P., Sperotto, A., Schotanus, H., Mandjes, M. and Meeuwissen, E. (2013) 'Flow-based detection of DNS tunnels', *International Conference on Autonomous Infrastructure, Management and Security*, pp.124–135, Springer Berlin Heidelberg.

Fachkha, C., Bou-Harb, E. and Debbabi, M. (2015) 'Inferring distributed reflection denial of service attacks from darknet', *Computer Communications*, Vol. 62, No. C, pp.59–71.

Fernandes, G., Carvalho, L.F., Rodrigues, J.J. and Proenca, M.L. (2016) 'Network anomaly detection using IP flows with principal component analysis and ant colony optimization', *Journal of Network and Computer Applications*, Vol. 64, No. C, pp.1–11.

Fernandes, G., Rodrigues, J. and Proenca, M.L. (2015) 'Autonomous profile-based anomaly detection system using principal component analysis and flow analysis', *Applied Soft Computing*, Vol. 34, No. C, pp.513–525.

Floyd, S. and Paxson, S. (2001) 'Difficulties in simulating the internet', *IEEE/ACM Transactions on Networking (TON)*, Vol. 9, No. 4, pp.392–403.

Gogoi, P., Bhuyan, M.H., Bhattacharyya, D.K. and Kalita, J.K. (2012) 'Packet and flow based network intrusion dataset', *International Conference on Contemporary Computing*, pp.322–334, Springer Berlin Heidelberg.

Gulisano, V., Callau-Zori, M., Fu, Z., Jimenez-Peris, R., Papatriantafilou, M. and Patino-Martinez, M. (2015) 'STONE: a streaming DDoS defense framework', *Expert Systems with Applications*, Vol. 42, No. 24, pp.9620–9633.

Hofstede, R., Celeda, P., Trammell, B., Drago, I., Sadre, R., Sperotto, A. and Pras, A. (2014) 'Flow monitoring explained: from packet capture to data analysis with netflow and IPFIX', *IEEE Communications Surveys & Tutorials*, Vol. 16, No. 4, pp.2037–2064.

Ippoliti, D and Zhou, X. (2014) 'Online adaptive anomaly detection for augmented network flows', *22nd International Symposium on Modelling, Analysis & Simulation of Computer and Telecommunication Systems (MASCOTS)*, pp.433–442, IEEE.

Jadidi, Z., Muthukkumarasamy, V. and Sithirasenan, E. (2013a) 'Metaheuristic algorithms based flow anomaly detector', *19th Asia-Pacific Conference on Communications (APCC)*, pp.717–722, IEEE.

Jadidi, Z., Muthukkumarasamy, V., Sithirasenan, E. and Sheikhan, M. (2013b) 'Flow-based anomaly detection using neural network optimized with gsa algorithm', *33rd International Conference on Distributed Computing Systems Workshops (ICDCSW)*, pp.76–81, IEEE.

Johnson, T. and Lazos, L. (2014) 'Network anomaly detection using autonomous system flow aggregates', *Global Communications Conference (GLOBECOM)*, pp.544–550, IEEE.

Karasaridis, A., Meier-Hellstern, K. and Hoeflin, D. (2006) 'Detection of DNS anomalies using flow data analysis', *Global Telecommunications Conference, 2006. GLOBECOM'06*, pp.1–6, IEEE.

Kumar, D.A. and Venugopalan, S.R. (2017) 'Intrusion detection by initial classification based on protocol type', *International Journal of Advanced Intelligence Paradigms*, Vol. 9, Nos. 2–3, pp.122–138.

Leach, J. and Tedesco, G. (2003) *Firestorm Network Intrusion Detection System*, Firestorm documentation [online] http://www.scaramanga.co.uk/firestorm/documentation.html (accessed 15 April 2016).

Li, B., Springer, J., Bebis, G. and Gunes, M. (2013) 'A survey of network flow applications', *Journal of Network and Computer Applications*, Vol. 36, No. 2, pp.567–581.

Li, Q., Wei, W., Tao, M. and Chen, Q. (2014) 'A DDOS defence scheme based on two-stage traffic flow control', *International Journal of Communication Networks and Distributed Systems*, Vol. 13, Nos. 3–4, pp.290–300.

Li, X. and Deng, Z. (2010) 'Mining frequent patterns from network flows for monitoring network', *Expert Systems with Applications*, Vol. 37, No. 12, pp.8850–8860.

Li, Y., Luo, X., Qian, Y. and Zhao, X. (2015) 'Network-wide traffic anomaly detection and localization based on robust multivariate probabilistic calibration model', *Mathematical Problems in Engineering*, Vol. 2015, Article ID. 923792, 26pp, doi:10.1155/2015/923792.

Liu, D., Lung, C., Lambadanì's, I. and Seddigh, N. (2013) 'Network traffic anomaly detection using clustering techniques and performance comparison', *26th Annual Canadian Conference on Electrical and Computer Engineering (CCECE)*, pp.1–4, IEEE.

Liu, W., Zheng, K., Wu, B., Wu, C. and Niu, X. (2016) 'Flow-based anomaly detection using access behavior profiling and time-sequenced relation mining', *KSII Transactions on Internet & Information Systems*, Vol. 10, No. 6, pp.2781–2800.

Marnerides, A., Schaeffer-Filho, A. and Mauthe, A. (2014) 'Traffic anomaly diagnosis in internet backbone networks: a survey', *Computer Networks*, Vol. 73, No. C, pp.224–243.

McAfee Labs Threats Reports (2015) August 2015 [online] http://www.mcafee.com/in/resources/reports/rpquarterly-threats-aug-2015.pdf (accessed 01February 2016).

Metaflows Security Systems [online] http://www.metaflows.com (accessed 10 April 2016).

Novakov, S., Lung, C.H., Lambadaris, I. and Seddigh, N. (2012) 'Combining statistical and spectral analysis techniques in network traffic anomaly detection', *Conference on Next Generation Networks and Services (NGNS)*, pp.94–101, IEEE.

OpenWIPS-ng [online] http://www.openwips-ng.org (accessed 10 April 2016).

Palmieri, F., Fiore, U. and Castiglione, A. (2014) 'A distributed approach to network anomaly detection based on independent component analysis', *Concurrency and Computation: Practice and Experience*, Vol. 26, No. 5, pp.1113–1129.

Paredes-Oliva, I., Barlet-Ros, P. and Dimitropoulos, X. (2013) 'FaRNet: fast recognition of high-dimensional patterns from big network traffic data', *Computer Networks*, Vol. 57, No. 18, pp.3897–3913.

Paredes-Oliva, I., Castell-Uroz, I., Barlet-Ros, P., Dimitropoulos, X. and Sole-Pareta, J. (2012) 'Practical anomaly detection based on classifying frequent traffic patterns', *Conference on Computer Communications Workshops (INFOCOMWKSHPS)*, pp.49–54, IEEE.

Paxson, V., Rothfuss, J. and Tierney, B. (2004) *Bro User Manual* [online] http://www.bro-ids.org (accessed 10 February 2016).

Pena, E., Barbon, S., Rodrigues, J., Lemes, M. and Junior, P. (2014) 'Anomaly detection using digital signature of network segment with adaptive ARIMA model and paraconsistent logic', *Symposium on Computers and Communication (ISCC)*, pp.1–6, IEEE.

Rehak, M., Pechoucek, M., Bartoš, K., Grill, M., Čeleda, P., Krmíček, V. et al. (2008) 'CAMNEP: an intrusion detection system for high-speed networks', *Progress in Informatics*, Vol. 5, No. 5, pp.65–74.

Rehak, M., Pechoucek, M., Grill, M., Stiborek, J., Bartoš, K. and Celeda, P. (2009) 'Adaptive multiagent system for network traffic monitoring', *IEEE Intelligent Systems*, Vol. 23, No. 3, pp.16–25.

Roy, A., Kar, P. and Misra, S. (2014) 'Detection of dumb nodes in a stationary wireless sensor network', *Annual India Conference (INDICON)*, pp.1–6, IEEE.

Roy, A., Kar, P., Misra, S. and Obaidat, M.S. (2017) 'D3: Distributed approach for the detection of dumb nodes in wireless sensor networks', *International Journal of Communication Systems*, Vol. 30, No. 1, p.e2913.

Salem, O., Naït-Abdesselam, F. and Mehaoua, A. (2012) 'Anomaly detection in network traffic using jensen-shannon divergence', *International Conference on Communications (ICC)*, pp.5200–5204, IEEE.

Sarmah, A. (2001) *Intrusion Detection Systems: Definition, Need and Challenges* [online] https://www.sans.org/reading-room/whitepapers/detection/intrusion-detection-systems-definition-challenges-343 (accessed 10 May 2016).

SecureWorks, Dell (2014) *Advanced Threat Protection with Dell SecureWorks Security Services* [online] https://www.secureworks.com/resources/sb-advanced-threat-protection (accessed 20 April 2016).

Sheikhan, M. and Jadidi, Z. (2014) 'Flow-based anomaly detection in high-speed links using modified gsa-optimized neural network', *Neural Computing and Applications*, Vol. 24, Nos. 3–4, pp.599–611.

Sperotto, A., Sadre, R., Van Vliet, F. and Pras, A. (2009) 'A labeled data set for flowbased intrusion detection', *International Workshop on IP Operations and Management*, pp.39–50, Springer Berlin Heidelberg.

Sperotto, A., Schaffrath, G., Sadre, R., Morariu, C. Pras, A. and Stiller, B. (2010) 'An overview of IP flow-based intrusion detection', *IEEE Communications Surveys & Tutorials*, Vol. 12, No. 3, pp.343–356.

Symantec Monitored and Managed IDS/IPS Services (2005) [online] http://eval.symantec.com/mktginfo/enterprise/fact_sheets/ent-factsheet_intrusion_detection_services_06-2005.en-us.pdf (accessed 10 April 2016).

Winter, P., Hermann, E. and Zeilinger, M. (2011) 'Inductive intrusion detection in flowbased network data using one-class support vector machines', *4th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*, pp.1–5, IEEE.

Yan, R. (2016) 'Combining Renyi entropy and EWMA to detect common attacks in network', *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 30, No. 10, p.1650021.

Yeung, D., Jin, S. and Wang, X. (2007) 'Covariance-matrix modeling and detecting various flooding attacks', *IEEE transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, Vol. 37, No. 2, pp.157–169.

Yu, S., Zhou, W., Jia, W., Guo, S., Xiang, Y. and Tang, F. (2012) 'Discriminating DDoS attacks from flash crowds using flow correlation coefficient', *IEEE transactions on Parallel and Distributed Systems*, Vol. 23, No. 6, pp.1073–1080.

Zang, X., Tangpong, A., Kesidis, G. and Miller, D. (2011) *Botnet Detection Through Fine Flow*, unpublished, Departments of CS&E and EE, The Pennsylvania State University, University Park, PA, Report No. CSE11-001.