ISSN NO. 2456-3129

IJEPAS

International Journal of Engineering, Pure and Applied Sciences,
Vol. 6, No. 2, June-2021

Forensic Network Traffic Analysis: An Overview

Ravinder Madhan, Sompal, Raman Kumar, Vikas Sheoran

Assistant Professor, Department of CSE, IEC University, H.P
Associate Professor, Department of CSE, IEC University, H.P
Assistant Professor, Department of Mechanical, IEC University, H.P
Assistant Professor, Department of Civil, IEC University, H.P
Email: ravindermadhan.cse@iecuniversity.com, sompal26@gmail.com, ramankumar.me@iecuniversity.com, sheoranvikas123@gmail.com

Abstract- Network forensics is a comparatively new field of forensic science. The growing popularity of the Internet in homes means that computing has become network-centric and data is now available outside of disk-based digital evidence. The real-world situations need a quick classification decision before the flow finishes, especially for security and network forensic purposes. Therefore, monitoring network traffic requires a real-time and continuous analysis, to collect valuable evidence such as instant evidences that might be missed with post-mortem analysis (dead forensics). Network traffic classification is the core component in evidence collection and analysis that uses filtered evidence and helps to reduce redundancy. However, most of the existing approaches that deal with collecting evidence from networks are based on post- mortem analysis. Therefore, this research investigates different classification techniques using Machine Learning (ML) algorithms, seeking to identify ways to improve classification methods from a forensic investigator standpoint because nature of information in a network is volatile and dynamic, some precious evidence might be missed.

Key words- Cyber security; Digital forensics; network traffic classification; digital evidence; network analysis.

1. INTRODUCTION

Network traffic analysis is the process of recording, reviewing and analyzing network traffic for the purpose of performance, security and/or general network operations and management. It is the process of using manual and automated techniques to review granular-level detail and statistics within network traffic. Network traffic analysis is primarily done to get in-depth insight into what type of traffic/network packets or data is flowing through a network. Typically, network traffic analysis is done through a network monitoring or network bandwidth monitoring software/application. The traffic statistics from network traffic analysis helps in:

- Download/upload speeds
- Type, size, origin and destination and content/data of packets
- Understanding and evaluating the network utilization

Network traffic analysis is also used by attackers/intruders to analyze network traffic patterns and identify any vulnerabilities or means to break in or retrieve sensitive data. Machine Learning (ML) techniques, which depend on analysis of application patterns that do not require payload inspection. Additionally, the increasing amount of traffic and transmission rates stimulate researchers to look for lightweight algorithms. And the persistence of application developers in inventing new ways to avoid filtering and detection mechanisms of traffic is another motivating factor.

2. WHAT IS DIGITAL FORENSICS?

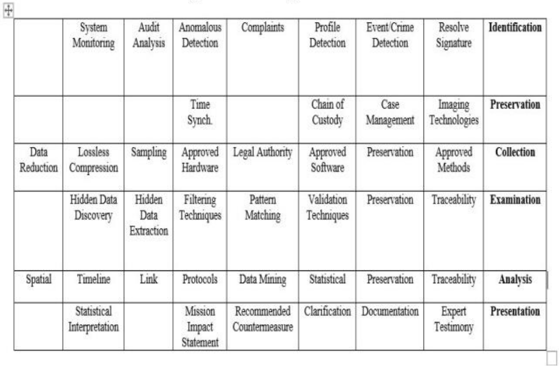
At the first Digital Forensic Research Workshop held in Utica, NY in 2001, the group created a consensus document which outlined the state of digital forensics at that time.The process of preservation, identification, extraction, and documentation of computer evidence which can be used by the court of law is known as digital forensics. It is a science of finding evidence from digital media like a computer, mobile phone, server, or network. It provides the forensic team with the best techniques and tools to solve complicated digital-related cases.Digital Forensics helps the forensic team to analyzes, inspect, identifies, and preserve the digital evidence residing on various types of electronic devices. Network forensics is a sub-branch of digital forensics Network forensics analyzes the network traffic and monitors data packets transferred over the internet for intrusion and malware detection. It involves collecting and recording data, analyzing the issue, determining the best troubleshooting response, and implementing it. Network forensics experts collect data from different websites and network equipment, including intrusion detection systems (IDS) and firewalls, to analyze network traffic data. Moreover, network forensics can also be used for monitoring, preventing, and analyzing potential attacks.

2.1 Background on Forensic Model

There are several models, which can be used for investigation in digital forensic science._**DFRWS Investigative Model was meant to be a generic "technology-independent" model, and in 2002** Mark Reith, Clint Carr, and Gregg Gunsch was inspired from DFRWS. Digital forensic analysis DFRWS model consists of sequential steps. The steps help the researchers to conceive of the situation to understand the direction of what they need to focus on. These
steps can be seen in **Table 1.**
Machine Learning (ML) has long been known as a powerful technique for data mining and knowledge discovery, which searches for and describes useful structural patterns in data. ML has a great range of applications, including in relation to search engines, medical diagnosis, text and handwriting recognition, image screening.

The location of the digital evidence and how to find the location this involves. For example, in network Forensics Intrusion Detection Systems (IDS) can be sources of digital evidence; IDS involves recognising and detecting unusual patterns of flow in the network traffic. The second step is preservation, which is a critical phase for increasing the possibility of a successful investigation. This process starts from acquiring, seizing, and preserving the evidence to create a digital image of the evidence and maintain the chain of custody. The following steps are collection, examination and analysis of the digital evidence to culminate in the final single presentation. Network forensics requires real-time collection of the digital evidence to avoid any missing critical information. However, most of the methods in the collection process are based on post- mortem analysis.

Table 1. **Digital Forensic Investigation Process**



	System Monitoring	Audit Analysis	Anomalous Detection	Complaints	Profile Detection	Event/Crime Detection	Resolve Signature	**Identification**
			Time Synch.		Chain of Custody	Case Management	Imaging Technologies	**Preservation**
Data Reduction	Lossless Compression	Sampling	Approved Hardware	Legal Authority	Approved Software	Preservation	Approved Methods	**Collection**
	Hidden Data Discovery	Hidden Data Extraction	Filtering Techniques	Pattern Matching	Validation Techniques	Preservation	Traceability	**Examination**
Spatial	Timeline	Link	Protocols	Data Mining	Statistical	Preservation	Traceability	**Analysis**
	Statistical Interpretation		Mission Impact Statement	Recommended Countermeasure	Clarification	Documentation	Expert Testimony	**Presentation**

The huge amount of collected data needs automated techniques to reduce redundancy, and consequently reduce the analysis time of the evidence. In addition, the massive amount of traffic needs a huge storage capacity, which costs a lot; this amount of data needs to be filtered to extract related data. However, analysis techniques based on pattern recognition using ML algorithms have proven promising results. Finally, the last stage of digital forensic investigation is presentation, which entails dealing with legal aspects of the case and presenting the investigation's findings in the court.

3. USING MLTECHNIQUES IN NETWORK TRAFFIC CLASSIFICATION

Network traffic Classification has generated great interest in the research community along with the Industrial field. In spite of the simplest and fastest of the previous classification technique (port-based) in monitoring and reporting activity of network traffic, the unreliability of port-based classification has been proved in several published works. Investigation of the accuracy of port-based classification demonstrated that the accuracy of this traditional technique, using the official IANAlist, is not better than 70% of bytes. Internet Assigned Number Authority (IANA) assigns port number for different applications. For example: TCP Port 80- HTTP, UDP Port 53- DNS & TCP Port 25-SMTP One study stated that the port base analysis method cannot classify 30- 70% of the internet traffic used in its work.
The drawbacks of port-based classification methods that some applications may not have their ports registered and applications may use ports other than its well-known ports. ML techniques have shown promising results in analysing network traffic based on the extracted features of the flow.

Machine Learning Concepts

ML has historically been known as a collection of powerful techniques for data mining and knowledge discovery, which search for and describe useful structural patterns in data. ML techniques are used in several applications such as medical diagnosis, search engines, marketing diagnosis, etc. A network traffic controller using ML techniques was proposed in 1990, aiming to maximize call completion in a circuit-switched telecommunications network. In 1994 ML was first utilized for Internet flow classification in the context of intrusion detection. In the early 1990s Shi noted that the distinctive characteristic of intelligence is that machines have the ability to learn form experience automatically. Additionally, in 2000 Witten and Frank spotted that "Things learn when they change their behavior in a way that makes them perform better in the future" This was the spark for a lot of research on applying ML techniques in network traffic classification.

Types of Machine Learning

To understand the pros and cons of each type of machine learning, we must first look at what kind of data they ingest. In ML, there are two kinds of data —

labeled data and unlabeled data. Labeled data has both the input and output parameters in a completely machine-readable pattern, but requires a lot of human labor to label the data, to begin with. Unlabeled data only has one or none of the parameters in a machine-readable form. This negates the need for human labor but requires more complex solutions.

According to Witten and Frank [6] there are four different types of Machine learning- supervised learning (classification), unsupervised learning (clustering), association, and numeric prediction. Supervised learning creates knowledge structure that supports the task classifying new instances into pre-defined examples, to build classifier rules to identify unknown flow. Unsupervised machine learning holds the advantage of being able to work with unlabeled data. Unsupervised learning automatically classifies network flow into groups (clusters) of instances that have the same properties without any kind of pre-guidance. Association learning explores the links among features. And in numeric prediction, the prediction result is the numeric volume, not a discrete class.

Two main processes are inherent in supervised classification—namely, the training process and then the final step, the testing process. The testing phase is the following process where the model (classifier) is used to identify new flow. Supervised techniques is suitable for identifying specific types of applications. The effectiveness of these techniques is subject to a training phase (training set), because these methods of classification focus on forming the relationships of input/output.

Supervised Classification

Supervised classification can be very effective and accurate in classifying satellite images and can be applied at the individual pixel level or to image objects (groups of adjacent, similar pixels). However, for the process to work effectively, the person processing the image needs to have a priori knowledge (field data, aerial photographs, or other knowledge) of where the classes of interest (e.g., land cover types) are located, or be able to identify them directly from the imagery. This method is often used with unsupervised classification in a process called hybrid classification. Many analysts use a combination of supervised and unsupervised classification processes to develop final output analysis and classified maps. The schemetic of the steps used in supervised classification are given below (Fig 2) Different ML algorithms that can be used in supervised classification. In supervised learning, algorithms learn from labeled data. Examples of these algorithms are the K-Nearest Neighbours (K-NN), Linear Discriminate Analysis (LDA) and Quadratic Discriminant Analysis (QDA) algorithms. The authors in[10], applied a statistical signature-based approach using these ML algorithms to classify IP traffic. Another study [11] used the supervised ML naive Bayes technique to classify network traffic and adopts the principle of class conditional independence from the Bayes Theorem. This means that the presence of one feature does not impact the presence of another in the probability of a given outcome, and each predictor has an equal effect on that result. There are three types of Naïve Bayes classifiers: Multinomial Naïve Bayes, Bernoulli Naïve Bayes, and Gaussian Naïve Bayes. This technique is primarily used in text classification, spam identification, and recommendation systems. In the training phase they used 248 full flow-based features. They were able to achieve 65% flow accuracy in their classification results with a simple Naive Bayes technique. The authors extended their previous study by using Naive Bayes and decision-tree algorithms. They affirmed once again that performance of classifiers is enhanced when they are trained using sub-flows with the same datasets. They also compared their results with the poor results of classification based on statistical features extracted from bidirectional flows.

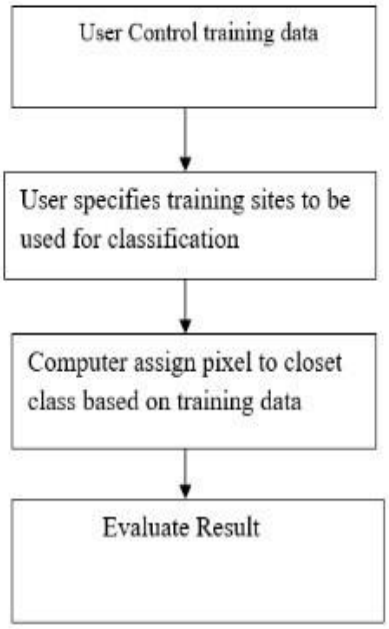

User Control training data
User specifies training sites to be used for classification
Computer assign pixel to closet class based on training data
Evaluate Result

Fig 2:- Steps of Supervised Classification

GAs are more robust algorithms that can be used for various optimization problems. These algorithms do not deviate easily in the presence of noise, unlike other AI algorithms. Another work [15] used Genetic Algorithm (GA) for feature selection, applied three different algorithms (Naive Bayesian classifier with Kernel Estimation (NBKE), Decision Tree J48, and the Reduced Error Pruning Tree (REPTree)) and compared their classification results. The accuracy result is based on overall results; it is noteworthy that high accuracy is provided by the first 10 packets used in the classification. This study raises the question: If different applications were used, how different would the results be?

Unsupervised Classification

Unlike supervised learning, unsupervised learning uses unlabeled data. From that data, it discovers patterns that help solve for clustering or association problems. This is particularly useful when subject matter experts are unsure of common properties within a data set. Common clustering algorithms are hierarchical, k-means, and Gaussian mixture models. Using this technique, the researchers grouped applications into several groups based on characteristics. Their technique can be taken as an initial step in identifying unknown network traffic. In general terms, an unsupervised classifier requires the following parameters to be specified by the user:

- Number of classes
- Number of bands
- Spectral distance or radius in spectral distance
- Spectral space distance parameters when merging clusters

Other researchers, such as [19], proposed a method to TCP flow by applying unsupervised ML (Simple K-Means algorithm). This classification method is based on the hypothesisthat assumes the classifier can always know the beginning of every flow. This is not the case in reality, as in real-world network traffic the start of the flow might be missed. The capability of the classifier decrease in situations different than studied conditions.

Semi-Supervised Classification

Semi-supervised learning occurs when only part of the given input data has been labeled. Unsupervised and semi-supervised learning can be more appealing alternatives as it can be time-consuming and costly to rely on domain expertise to label data appropriately for supervised learning.One way to do semi-supervised learning is to combine clustering and classification algorithms. Clustering algorithms are unsupervised machine learning techniques that group data together based on their similarities. The clustering model will help us find the most relevant samples in our data set. We can then label those and use them to train our supervised machine learning model for the classification task. The initial results with a semi-supervised approach using the K-Means clustering algorithm is promising.

4 RESEARCH CHALLENGES

An essential aim for a forensic investigator is to determine how a crime was undertaken – in order to identify the actors involved, how the crime unfolded and to develop a chronology of the incident. The tools developed are for the investigator. The information provided by the tools has to be at very abstract level leaving behind the technical intricacies. The Quantity Problem in Network Forensics is that the amount of data to be analysed can be very large. Also, it is notable that most studies applied ML techniques using features extracted from full flows, whereas the results of classifying using sub-flows outperformed the results of studies using full flows. Indeed, for security analysis purposes the speed in identifying network traffic is also required as the accuracy of the results. The problem of classifying using full flows is that if the beginning of the flow is missed, performance of the classifier diminishes. In addition, losing the beginning of the packet is a very common scenario in realworld network traffic.

5 THE PROPOSED TECHNIQUE

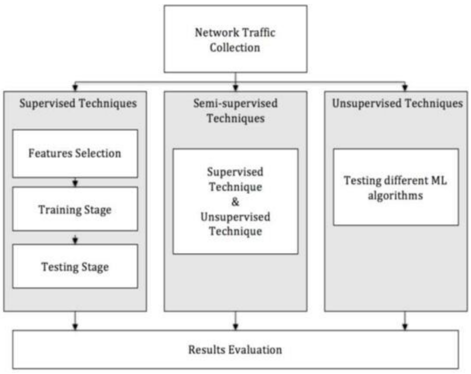
The idea behind this research is that we can evaluate the extent of their effectiveness in coping with real-world situations—situations that require both the prompt recognition of application type in the network before the flow ends, and the making of quick decisions. Such timeliness is critical, as many criminal incidents could be detected and prevented before they can inflict extensive damage.

The first step involves capturing network traffic to create an appropriate dataset for our investigation. The second step consists of extracting features, such as the duration and length of the packet and inter-packet arrival times, from the sub-flow. We chose to work with features extracted from sub-flow because we believe this is the fastest way to analysis network traffic instead of using full-flow especially for security and forensic aspects. The extracted features will be used to train the classifier in the supervised techniques stage, with the endpoint of analysing network traffic smoothly.

In the next stage, the objective will be to enhance the performance of the clarifier by combining supervised and unsupervised learning techniques. We will then evaluate the results of using both types of technique, to determine any differences in performance. In the third stage, we will investigate the capability of unsupervised techniques to work independently. Through this work, we will look to find a way of improving the performance of classification techniques, without relying solely on the use of supervised techniques. Additionally, we look also to emphasize the importance of apply unsupervised techniques in network traffic analysis and investigate the ability of these techniques to distinguish new and unseen applications that could be malicious in nature.

Fig 3: The Workflow of Forensic Network Traffic Analysis

6.


Network Traffic Collection
Supervised Techniques
Features Selection
Training Stage
Testing Stage
Semi-supervised Techniques
Supervised Technique & Unsupervised Technique
Unsupervised Techniques
Testing different ML algorithms
Results Evaluation

CONCLUSION

Network forensics ensures a faster incident response to an attack. It provides the ability to investigate the attacks by tracing the attack back to the source and discovering the nature of the attacker if it is a person, host or a network. In addition, network forensics provides methods to predict future attacks by correlating attack patterns from previous records of intrusion traffic data. For computer security, it has been proposed that forensic network analysis introduce investigative capabilities into current networks. This facilitates the presentation of admissible evidence in a court of law. In fact, to build a clear and strong case lawyers need more corroborating evidence, which calls for real-time collection. Therefore, this research aims to identify methods that can improve classification techniques by using ML algorithms, and create a balance among accuracy, efficiency, and cost, because a large amount of network traffic is still unclassified.

REFERENCES

[1]. Digital Forensics Research Workshop. (2001, November). A road map for digital forensics research 2001. Retrieved from http://www.dfrws.org

[2]. W. Wang and T. Daniels. “Building Evidence Graphs for Network Forensics Analysis,” in Proceedings of the 21st Annual Computer Security Applications Conference (ACSAC 2005). September 2005.

[3]. A. Madhukar and C. Williamson, “A longitudinal study of P2P traffic classification,” in 14th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, September 2006.

[4]. A. W. Moore and K. Papagiannaki, “Toward the accurate identification of network applications,” in Passive and Active Measurement, Boston, MA, March 2005, pp. 41–54.

[5]. Z. Shi, Principles of Machine Learning. International Academic Publishers, 1992.

[6]. I. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, 2nd ed., Morgan Kaufmann Publishers, 2005.

[7]. B. Silver, “Netman: A learning network traffic controller,” in Proceedings of the Third International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, Association for Computing Machinery,1990.

[8]. J. Frank, “Machine learning and intrusion detection: Current and future directions,” in Proceedings of the National 17th Computer Security Conference, Washington, D.C., October 1994.

[9]. Y. Reich and J. S. Fenves, “The formation and use of abstract concepts in design,” in Concept Formation: Knowledge and Experience in Unsupervised Learning, D. H. Fisher and M. J.Pazzani, Eds. Morgan Kaufmann, 1991.

[10]. Kumar Sharma, A., & Sharma, S. K. Vibration Computational of Visco-Elastic Plate with Sinusoidal Thickness Variation and Linearly Thermal effect in 2D. Journal of Advanced Research in Applied Mechanics & Computational Fluid Dynamics, 1(1), 2014, 46-54.

[11]. A. Moore and D. Zuev, “Internet traffic classification using Bayesian analysis techniques,” in ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS)2005, Banff, Alberta, Canada, June 2005.

[12]. Sharma, S. K., & Sharma, A. K. Mechanical Vibration of Orthotropic Rectangular Plate with 2D Linearly Varying Thickness and Thermal Effect. International Journal of Research in Advent Technology, 2(6), 2014, 184-190.

[13]. T. Nguyen and G. Armitage, “Training on multiple sub-flows to optimise the use of machine-learning classifiers in real-world IP networks,” in Proceedings of the IEEE 31st Conference on Local Computer Networks, Tampa, FL, pp. 369–376, November 2006.

[14]. T. Nguyen and G. Armitage, “Synthetic sub-flow pairs for timely and stable IP traffic identification,” in Proceedings of the Australian Telecommunication Networks and Application Conference, Melbourne, Australia, December 2006.

[15]. J. Park, H.-R. Tyan, and K. C.-C. J. Kuo, “GA-based Internet traffic classification technique for QoS provisioning,” in Proceedings of the 2006 International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Pasadena.