Detection of Crime Patterns in Digital Forensic Investigation to Trace the Adversaries

Rida Zahra¹, Muhammad Ilyas², Ahmad Bilal³, Qaiser Abbas⁴, Wakeel Ahmad⁵, Syed M Adnan⁶

¹Department of Computer Science, University of Wah, Wah, Pakistan

^{2,4}Department of Computer Science and IT, University of Sargodha, Sargodha, Pakistan

³School of Electrical Engineering and Computer Science, NUST, Islamabad, Pakistan

^{5,6}Department of Computer Science, University of Engineering and Technology, Taxila, Pakistan
Email: muhammad.ilyas@uos.edu.pk.

ABSTRACT

The use of the internet has increased significantly over the past couple of years. Access to the internet has become so common that a person without computer knowledge can also use this facility easily. This ease of availability has provided a lot of benefits to society but on the other hand misuse of the internet for personal or corporate benefits is also increasing. To prosecute cybercriminals and make some lawful checks on everyone's digital activities, digital forensic science comes into the light. In this context, we developed a new framework that improves the digital forensic investigation process. This research paper proposes a method in which we can identify the illegal activities and trace the adversaries. We capture the TCP (Transmission Control Protocol) packets from the servers and workstations. This data collected from the TCP log is stored in the database and preprocessed to eliminate redundant data. Furthermore, the database also contains past data. The proposed framework has three major processes collection of TCP packets, storing and preprocessing of collected data in a database, and mining of the pattern through a digital forensic anomaly collection algorithm. For the evaluation of our proposed framework, we have developed a java based application. The results are shown in the form of reports and tables.

KEYWORDS: Cyber-crime, Digital forensic, Data mining, Crime pattern, Cybersecurity

1. INTRODUCTION

Cybersecurity has become one of the most important things in today's computer domain. Many kinds of computers and internet crime are discovered including stealing personal information, fraud, embezzlement, harassment, white-collar crime, data diddling, hacking, social media crimes, salami attack, zero-day attack, wardriving attack, phishing attack, spoofing attack, and many other [1]. Organizations save their precious data over the computer, share their data over the internet. As shown in figure 1, the rate of cybercrime is increasing in the past few years many popular organizations have become a victim of cybercrime [1][2].

Most of the cases are not being reported

[3]. The reason that a low number of cases are reported to law enforcement is that most companies or business figments are more conscious about the publicity of their company rather than the loss they face [3]. We need a better secure technology solution to overcome this and a better cure to tackle this disease to get its root and destroy it. Better training of new technology tools and advanced methods to investigate is needed so that if any unexpected thing happens in terms of the digital domain we can handle it [4]. For this purpose, digital forensic science is getting the eyes of researchers in recent years. The rest of the paper compromise the following sections. Section 2 describes the Background. Section 3 discusses related work. The proposed framework and statistical evaluation of the proposed framework

are given in Sections 4 and 5 respectively. Results and discussions are explained in Section 6. Section 7 comprises the conclusion and future work.

Figure 1: Increase in Crime Rate [1]

2. BACKGROUND

In the past year, a new way of crime is becoming more common, which is through digital media and within the cyber domain. To investigate these types of crime a new domain comes in front known as “Digital Forensic”. Digital forensic, also known as digital forensic science is emerging from forensic science. Digital forensic science began as the result of cybersecurity breaches to trace the cybercriminal and confront them in court based on collected digital evidence which is collected from the targeted digital devices. It deals with the recovery and material (commonly known as digital evidence) found in digital devices [5]. Digital forensic science is defined as; the use of the scientifically driven method of collecting, preserving, analyzing, validating, and documentation the digital evidence taken from digital sources for the construction of events that are suspected [4].

French criminalist Edmond Locard Presented the principle of exchange which states: “When a person commits a crime something is always left at the scene of the crime that was not present when the person arrived [6]. So, it is clear by the above statement that something minor or major is always left behind when a crime is committed. The main purpose of the digital investigation process is to find digital evidence to present them in court.

Digital evidence is the most important part of the digital Investigation process because

the whole crime case hinges on it [7]. Digital evidence is the entity that provides the traces to tackle the illegal activity, and it is the only thing that can be provided to the court as proof. The evidence found in physical forensic science is permanent by its nature [7]. For example, if there is a murder case investigators could find a weapon, fingerprints, or any other clue which is preserved and cannot be damaged or remove easily. This is not the case when we talk about digital evidence. Digital evidence is very sensitive to handle and volatile by nature. It can easily change and alter or even damaged permanently.

Data mining combines itself different discipline which are databases systems, machine learning, statistics, visualization, and information science. Data mining varies from one discipline to the other based on the data to be mined. Different discipline as mention above contributes their part in the making of a data mining expert system [8]. So, the data mining system needs to be classified differently, so users can choose the right one per their specification.

3. RELATED WORK

Ankur Kumar et al. [9] presented a digital forensic investigation model. As digital crime is increasing drastically all over the world in the form of unauthorized access through a new type of computer and cyber-attacks, so it becomes very important that how security policies are going to design and deploy to overcome such intrusions. So, to make malicious users, intruders or attackers 3R model should be adopted by the organizations for keeping an eye on these kinds of situations. The 3R model cover three aspects which are Resistance, Recognition, and Recovery. The first R (Resistance) is for a preventive measure which means the system should be able to prevent intruders, the second R (Recognition) is for identifying the illegal intrusion to the system and the third R (Recovery) is used to lose the overall impact of the incident and to bring back the system to normal. This 3R model also covers digital forensic. This 3R model also describes the forensic investigation requirements. In the DFIDM digital forensic investigation integrated into the best policies practices, procedures, techniques, and training to assure the security of information.

Yusoff et al. [10] have identified the most common phrases in the digital investigation process from the different models of digital

investigations. Those models are Computer Forensic Investigation Process which was presented in 1995, the DFRWS Investigation model and Scientific Crime Scene Model in 2001, two models in 2003 which are Abstract Digital Investigation Model and the End to End Digital Investigation Model. A more improved model for digital forensic science was presented by researchers, these are Enhance Digital Investigation Model, Extended Model for Cybercrime Investigation, A Hierarchical: Objective-Based Framework for the Digital Investigation, Computer Forensic Field Triage Process Model, Framework for Digital Forensic Investigation, Dual Data Analysis Process, Common Process Model for Incident and Computer Forensic and Digital Forensic Model Based on Malaysian Investigation Process. These all mentioned models and frameworks have general five phases which are presented in a different context. These phases are generally named pre-process, acquisition and preservation, analysis, presentation, and post-process.

Sharma et al. [11] proposed a framework based on a physical crime scene procedure in which each device is considered as a separate crime scene. This model's investigation includes the system's preservation, extraction of digital evidence, and reconstruction of the digital event. A hypothesis-based analysis is included in this model. The event-based framework works on the principle of physical crime scene forensic investigation to develop a hypothesis for detecting the previous incident using the object. The authors have proposed a new tool that combines digital forensic processes and crime data mining to form a hybrid application [12]. The purpose of this combination is to find the motive and pattern of cyber-attacks and the frequency of attacks. Hence, system vulnerability is minimized by this tool. A crime data mining algorithm is used which is based on the Apriori algorithm. In this algorithm, the first step is to identify the items from the report. There are item sets $\{I=I_1, I_2, I_3, \dots, I_m\}$ and there are a set of actions to perform on these items. Then frequent item sets are mined using Apriori, then association rules are made to check on the test dataset. Then SQL queries are set per the rules. A misuse-based IDS was presented by Mansour Sheikhan and Zahra Jadidi [13] which is a hybrid technique. It uses the structure of two techniques which are an association rule mining algorithm and a connections model. The main

thought behind merging this combo was to use the advantages of both pieces of knowledge-based and machine learning techniques. This is done to trace different attacks. As the result of association rule mining many rules are generated, to decrease this computational load the inputs of rule mining are based on the result of feature relevance analysis. After the implementation of this hybrid model, the results show that the system can report some critical attack categories, and it can give better results for the identification of attacks and especially for R2L, U2R, classes.

Ronald et al. [14] in their work presented how to increase digital investigator availability through efficient workflow management and automation. The main purpose in their work is to reduce the cost and providing forensic investigator more time to spend on quality case report as the capacities in digital storage is increasing, it has a significant time impact on digital forensic laboratories. The digital forensic investigation preparation is taking more time than detailed evidence analysis and reporting.

In this context, they have presented a workflow for the management automation for handling common digital forensic tools. This would provide more efficient use of hardware and software, and it also reduces the time which digital forensic experts waste conducting a different investigation process. The presented work is tested in a real-world scenario. The presented framework is composed of three workflows. The first framework is image creation; it helps in controlling the image creation process. The registered user provides information about the committed crime and its nature and is asked to select the preparation steps. After the successful completion of the acquisition, the preparation task will be started without losing any time.

The second component Queue Server is for the implementation of the queue server which is for controlling third-party software. It will monitor a queue folder and checks on the time intervals for the job in queue for processing. Here job refers to the file store in the queue directory containing information.

The third component is for cleanup and archiving that will run to clean up and archive closed cases. It will read the configuration file, make a copy of the case registration and check each folder in the directories.

Figure 2: Digital Forensic Anomaly Collector Framework

4. PROPOSED FRAMEWORK

A framework, DFAC, has been proposed for collecting the anomalies from network behavior. Framework DFAC is shown in figure 2 below.

4.1. Collecting TCP (Transmission Control Protocol) Data.

In the first phase of our framework TCP logs are collected from the workstations or server. These logs contain the basic information about source and destination and data travel through the network. TCP works on the transport layer and is called transport layer protocol [10].

Transport layer is the layer of OSI (open system interconnection model) whose basic responsibility is to maintain end to end communication over internet [10], this layer also ensures the reliability of the packet and handles errors and data lose if any. TCP is used by the application to make sure the guaranteed delivery of data packets. When data is transferred from source to destination the data packet contain a MAC (Media Access Control) header, which is at the beginning of the network packet to turn it into a frame, IP header, TCP header and data.

4.2. Saving Collected Data in Databases

The second phase of our above shown framework is to save the collected logs in the databases. Collected data should be saved as it is collected in its original form and further processing is done separately. If the processed data is lost or damaged later, we have the original data in databases. The test data is saved here, when there are no incoming server's TCP logs.

4.3. Data Pre-processing

Data pre-processing is the important step in database management. When data is collected in a large amount from some place there are chances of redundant and duplicate data. Data preprocessing makes the data stable. If that data is included in the experiment it may not give the accurate result as wanted. For this reason, collected data is processed to remove any duplicate or redundant entries from collected logs.

4.4. Extracting and Analyzing Digital Evidences

After the preprocessing of data, the next step is to extract the digital evidence. Those entries which provide a strong connection with the case are the digital evidence and this is done with the help of attributes. The best attributes in

this context are those which closely relate patterns of all activities to the phenomenon under consideration. This extracted digital evidences analyzed per the current situation and further improvement is made.

4.5. Detection of Pattern Based on Attribute

In this part of the framework, the patterns are further separated purely based on attributes. Sometimes all attributes are not necessary for including in a pattern detection. Different attributes give a different type of information in different situations. Two types of the threshold for selection of best attribute are set for this purpose. Threshold1 will identify that an attribute's value is consistent or not. Our second threshold will check the occurrence of each attribute with the other attribute of the dataset. So here patterns are separated based on those attributes which give pure information to help in the current situation.

5. STATISTICAL EVALUATION OF PROPOSED FRAMEWORK

The statistical evaluation of the proposed framework is done in WEKA (Waikato Environment for Knowledge Analysis) [15]. It is used for the selection of the best attributes which can help in classifying the data more efficiently. The learning rule classifiers used are: Conjunctive rule learner, ZeroR Learner and Ridor Learner.

Conjunctive Rule learner can predict for both numeric and nominal values. The rule consists of antecedents "AND" and the class value. If the test instant is not covered by the adopted rule it is covered by the rule in training dataset. The conjunctive rule learner selects antecedent by using its information gain value and prunes the generated rule by using Reduced Error Pruning (REP) or simple pre-pruning [15].

ZeroR Classifier is the meekest method that depends on the target and ignores all predictors. Its target is to simply predict the majority class category. It is useful when defining a starting point for other classification methods [16].

Ridor is known as Ripple Down Rule Learner. It works by generating first a default rule and the rules of the default rule with minimal error weight. Then it generates the best rules for each individual rule and iterate until pure. The Set of rules are called Exceptions that predict classes [17].

5.1. Data Collection

The dataset which we have used to check our prototype is NSL-KDD which is the updated version of KDD cup 99 datasets [18]. This dataset is composed training and test data. The training dataset has 21 different attacks which are 37 % of test dataset [18]. The Cross-validation folds' value is set to 10 in this. The attack types are grouped into four categories: DoS, U2R, Probe and R2L. The number of instances in training and test dataset are shown in the table 1 below:

Table 1. Number of Instances in Training and Test Datasets

Categories	Instances	
	Training data	Testing data
Normal	67343	9711
DOS	45927	7456
Probe	15656	2421
U2R	52	200
R2L	995	2756

6. RESULTS AND DISCUSSION

The result of classifiers is shown below in table 2:

The first classifier Conjunctive Rule Learner is correctly Classifying 23692 instances, acquiring an accuracy rate of 94.0457%. The incorrectly classified instances are 1500. TP (True Positive) rate for class normal is 0.921 and TP rate for the class anomaly is 0.963. The FP (False Positive) rate for class normal is 0.037 and FP rate for anomaly class is 0.079. The detailed output is shown in table 3.

The second classifier ZeroR Learner is correctly classifying 13449 instances. The accuracy rate of this classifier is 53.386 %. The incorrectly classified instances are 11743. TP (True Positive) rate for class normal is 1 and TP rate for the class anomaly is 0. The FP (False Positive) rate for class normal is 1 and FP rate for anomaly class is 0. Time taken by this classifier is 5.95 seconds. The detailed result is shown in table 4.

The third classifier Ridor Learner is correctly classifying 25091 instances. The accuracy rate is 99.5991 %. Incorrectly classified instances are 101. TP rate for normal and anomaly class is 0.999, 0.992 respectively. FP rate for normal and anomaly class is 0.008,

0.001 respectively. The output for Ridor Rule Learner is shown in table 5:

The top two best classifiers are Ridor Learner and Conjunctive Rule Learner. As compared to ZeroR classifier, both are achieving an accuracy of 99.59% and 94.04% respectively. Even though ZeroR classifier is producing better results (in percentage), but it is taking too much time, almost two hundred percent more execution time. So, on such values, we will say that Conjunctive Rule Learner is a better classifier as compared to the other two classifiers. The classifiers may vary in result if some other dataset is used. Different classifiers give different results per the dataset. In this work, only one type of data set is used, so in this context, the Conjunctive Rule Learner classifier

is performing well.

We have also compared our results with other studies too. Seyedamin et al. [19] also used the classifier Conjunctive Rule Learner along with many other classifiers for the prediction of heart disease. They achieved an accuracy of 69.96%, whereas, in our studies, it's 94.04%. Pradeep [20] used multiple classifiers in his studies to predict software faults. He used both Ridor and Conjunctive Rule Learner classifiers and achieved 91.21% and 92.19% respectively as compared to our results 99.59% and 94.04% respectively. In both classifiers, we have achieved better results.

Table 2. Classifiers Results

Rule Learner	Correctly Classified Instances	TP Rate Normal	TP Rate Anomaly	FP Rate Normal	FP Rate Anomaly	Time (seconds)
Conjunctive Rule Learner	23692	0.921	0.963	0.037	0.079	2.27
ZeroR Learner	13449	1	0	1	0	5.95
Ridor Learner	25091	0.999	0.992	0.008	0.001	49.92

Table 3. Output of Conjunctive Rule Learner

Normal	Anomaly	Total Number of Instances	Correctly Classified Instances	Incorrectly Classified Instances	Time taken (Sec)
0.085998	0.914002	25192	23692 (94.04 %)	1500 (5.954 %)	2.27

Table 4. Output of ZeroR Learner

ZeroR predicts class value	Total Number of Instances	Correctly Classified Instances	Incorrectly Classified Instances	Time taken
Normal	25192	13449 (53.386 %)	11743 (46.614 %)	5.95 seconds

Table 5. Output of Ridor Learner

Total number of rules	Total Number of Instances	Correctly Classified Instances	Incorrectly Classified Instances	Time taken
15	25192	25091 (99.5991%)	101 (0.4009%)	199.14 seconds

7. CONCLUSION AND FUTURE WORK

In this paper, a framework is proposed regarding Digital Forensic Investigation Process. The main idea is to detect crime patterns by capturing internet behavior. We have done this by using data mining techniques and classifier evaluation in WEKA. This framework is consisting of three processes. The first process is to collect TCP logs and storing them in databases. The second process is to apply the algorithm and find the digital crime pattern. The third is to generate a result. Three classifiers ZeroR Learner, Ridor Learner and Conjunctive Rule Learner were used to evaluate the performance of the framework. Conjunctive Rule Learner classifier outperformed as compared to other classifiers and achieve an accuracy of 94.0457%.

In the future, we would implement an algorithm and build a prototype of our proposed work. We would also try to take the real-world dataset on run time to run the algorithm and to test results on different other classifiers.

REFERENCES

- [1] J. Clement. "IC3: total damage caused by reported cybercrime 2001-2019," www.statista.com/statistics/267132/total-damage-caused-by-by-cyber-crime-in-the-us/. (Accessed Feb. 28, 2021).
- [2] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Oxford, England: Morgan Kaufmann, 2012.
- [3] Studies of IT practitioners. "The Impact of Cyber Crime on Business," www.ponemon.org/local/upload/file/Impact_of_Cybercrime_on_Business_FINAL.pdf (Accessed Feb. 28, 2021).
- [4] G. Palmer, "A Road Map for Digital Forensic Research," New York, Tech. Rep. DTR-T001-01, 2001.
- [5] EC-council. "How Well Do You Know Digital Forensics?," www.eccouncil.org/what-is-digital-forensics/. (Accessed Mar. 01, 2021).
- [6] Edmond Locard, "A Simple guide to trace Forensic Evidence," www.forensicsciencesimplified.org/trace/TraceEvidence.pdf (Accessed Mar. 01, 2021).
- [7] M. Ilyas and R. Zahra, "Case Oriented Digital Evidence Similarity Framework", *Journal of Applied Environmental and Biological Sciences*, vol. 4, no. 9S, pp. 205-209, 2014.
- [8] J. Talar, "Data mining methods-application in metallurgy", *Archives of Metallurgy and Materials*, vol. 52, no. 2, pp. 239-250, 2007.
- [9] A. K. Shrivastava, et al., "Digital forensic investigation development model," In *Proc. IEEE 5th International Conference and Computational Intelligence and Communication Networks*, 2013, pp. 532-535.
- [10] Y. Yusoff, R. Ismail, and Z. Hassan, "Common phases of computer forensics investigation models", *International Journal of Computer Science & Information Technology*, vol. 3, no. 3, pp. 17-31, 2013.
- [11] B. K. Sharma, M. A. Joseph, B. Jacob and B. Miranda, "Emerging trends in Digital Forensic and Cyber security- An Overview," In *Proc. 2019 Sixth HCT Information Technology Trends (ITT)*, Ras Al Khaimah, United Arab Emirates, 2019, pp. 309-313.
- [12] K. k. Sindhu and B. B. Meshram, "Digital forensics and cybercrime datamining", *Journal of Information Security*, vol. 3, no. 3, pp. 196, 2012.
- [13] M. Sheikhan and Z. Jadidi, "Misuse detection using hybrid of association rule mining and connectionist modeling", *World Applied Sciences Journal*, vol. 7, pp. 31-37, 2009.
- [14] R. I. D. Braekt, N. Le-Khac, J. Farina, M. Scanlon and T. Kechadi, "Increasing digital investigator availability through efficient workflow management and automation," In *Proc. 2016 4th International Symposium on Digital Forensic and Security*, Little Rock, AR, USA, 2016, pp. 68-73.

- [15] Weka. "Conjunctive Rule Learner," <http://weka.sourceforge.net/doc/packages/conjunctiveRule/weka/classifiers/rules/ConjunctiveRule.html> (Accessed Mar. 01, 2021).
- [16] H. Saed. "ZeroR Classifier," www.saedsayad.com/zeror.html (Accessed Mar. 01, 2021).
- [17] C. Ridor. "Ridor Rule Learner," www.eecs.yorku.ca/tdb/_doc.php/userg/sw/weka/doc/weka/classifiers/rules/Ridor.html (Accessed Mar. 01, 2021).
- [18] S. Revathi, and A. Malathi, "A detailed analysis on NSL-KDD dataset using various machine learning techniques for intrusion detection", International Journal of Engineering Research & Technology, vol. 2, no. 12, pp. 1848-1853, 2013.
- [19] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia and J. Gutierrez, "A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease," 2017 IEEE Symposium on Computers and Communications (ISCC), 2017, pp. 204-207, doi: 10.1109/ISCC.2017.8024530.
- [20] P. Singh, "Comprehensive model for software fault prediction," 2017 International Conference on Inventive Computing and Informatics (ICICI), 2017, pp. 1103-1108, doi: 10.1109/ICICI.2017.8365311.