Digital Forensic Process in Cyber Crime Data Mining

Jagadeesha.G.M, Kotrappa.Sirbi, Veeragangadhara Swamy.T.M

***Abstract:** With enormous developments in communication and information technology, the rate of crimes increasing exponentially using technology. Data Scientists are in very much need of addressing the security issues with cyber-crime. The crimes will be committed with the help of digital devices, Data Scientists need to follow some actual outlines and methodologies to mend the raw information for investigation will be used as important evidences. In this paper ,Crimes at cyber operations will be dealt with forensic investigation, phases at Storage media, analysis of Unseen evidence at file organization, investigation methods for Network, investigation Methods for Memory & data mining for Cyber Forensic. Mining Algorithm Apriori is used to find the cyber-attack pattern and also system uses the tools like Winpcap, jpcap for background process. Integrating all the techniques discussed above will be used to predict the cyber-attack suspect.*

***Keywords:** Cyber Crime, Data Mining, Data Science, Digital Forensic, Network Investigation.*

I. INTRODUCTION

Data science is the science of learning scientific knowledge. Data science has received widespread attention in tutorial and industrial circles. Presently, there are many viewpoints regarding the data science definition. Many disciplines use knowledge technology to take care of scientific knowledge from their various areas. Many disciplines use knowledge technology or data science to take care of scientific knowledge from their various areas.

This work addresses the issues related to the cyber-attacks by utilizing methodologies of File system investigation, analysis of Hidden evidence at file system, methods for Network investigation, Memory investigation Methods and Cyber Forensic.

- From the deleted files, File system Investigation module finds the evidences, empty spaces, slack files etc. This method discovers every files & directories available with the drive, next identifies files which are removed from the memory. File system investigation gathers proofs at the files system, searches for the data in the slack space, empty space, and also at deleted space. Extract the hidden or deleted data. Further the data is processed for forensic examination.
- Network investigation methods are prepared for evidence gathering from the traffic monitoring tool. live forensic information will be taken from packet analyzer, concerning associate attack. system uses the Winpcap and Jpcap tools to monitor this process. These tools monitor the networking details of the system.
- Memory investigation Methods uses Wmic tool to collect the previous browsing information with process which are active at the computer. This methodology includes list of all the running processes in the computer, all the ports which are active, all the running different services on the computer, distribution of Load and users login sessions information.
- Cyber Forensic data mining uses Apriori algorithm. It uses mining algorithm called Apriori which is a mixture of test data and training data. Then relate test data with training data by comparing all the collected information. After comparison classify the repeated sets of items. After all these steps apply rule of association. Lastly, results of prediction are found.

II. PROBLEM STATEMENT AND MOTIVATION

Nowadays Tools for forensic were not perfect for the following events:

- To find data correlations - **Association**
- Identifying and arranging the data into groups of similarity, on the basis of collected information - **Classification**
- Determining earlier unknown information or the data which are not noticed will be recognized and presented as a cluster for visualization - **Clustering**
- Discovering patterns and information that will cause affordable predictions - **Forecasting**:

Combining and analyzing above said tasks are not up to the perfection. There is a need of a new tool, which will give more accuracy than earlier tools.

III. CONTRIBUTION

In this work, following 4 contributions were made

- File system Investigation at the Storage media, Hidden data analysis in the file system.
- Network investigation methods is equipped with a traffic monitoring tool for data/evidence collection
- Memory investigation Methods consist of the browser history and active process in the system.
- Cyber Forensic data mining uses Apriori algorithm to find pattern and analysis of crime suspect.

We are proposing a new tool which is the combination of digital forensic investigation and crime data mining.

Revised Manuscript Received on December 22, 2018.

Jagadeesha.G.M Rao Bahadur Y Mahabeshwarappa Engineering College, VTU, Belagavi

Kotrappa.Sirbi, KLE's Dr.M S Sheshgiri College of Engineering, VTU, Belagavi

Veeragangadhara Swamy.T.M, Rao Bahadur Y Mahabeshwarappa Engineering College, VTU, Belagavi

IV. LITERATURE SURVEY

- Sindhu. K. K. and Dr. B. B.Meshram[1] proposed a paper on file system about evidences which are hidden and recovery of hidden files. They explains the concept of network packets and those detail investigation like live packet analysis and also it addresses mining algorithms which are concerned to analysis of crime occorances.
- Prashant K. Khobragade and Latesh G. Malik[2] proposed a paper for log analysis through which crime occorance analysis can be done using digital forensic tool.Analysis of information at physical and logical memory will give more assistance for crime investigation.
- Sonal Honale and Jayshree Borkar[3] proposed a paper which uses mining algorithm like k means and apriory to identify the crime pattern and also this work proposes the methodology to collect the evidences from the network and file system with the identification of denial of service and sql injection attacks.
- Nikhil Kumar Singh, Deepak Singh Tomar and Bhola Nath Roy [4] implemented a work about identifying crime using relational algebra with the help of logs which are generated from different sources.It will use decision tree to classify the malicious user among the user group.
- V Anitha and Dr.P.Isakki[5] proposed a paper on mining techniques which uses server logs to analyze and predict the unusual user behavior based on the pattern identification and visualization.It uses student and faculty related websites for prediction.
- Raburu George, Omollo Richard and Okumu Danie proposed a paper[6],which demonstrates distribution of different files at user cluster drive using classification of attributes. The file existence at hard drive, which are more than once and it access meta data of the file, which reflects the behavior of a particular of a person. By this method ownership will be identified by using user profiles at windows.
- Simon L. Garfinkel [7] proposes a paper for information processing for recent abstractions and predicts the digital forensic impending crises of current trends observed by observers and it gives a proper solution to address the problem
- Mandeep Kaur,Navreet Kaur, Suman Khurana[8] are proposes a paper on tools available for analysis of information,which helps in identifying the cyber forensic crimes. This paper gives a comparison based on efficiency of different tools.

V. METHODOLOGY

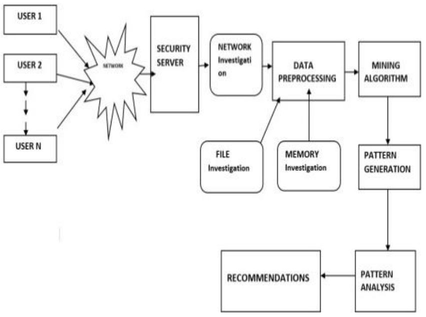


Figure 1: Block diagram of a Cyber Crime Data Mining

Above block diagram describes the proposed methodology and different modules, which contributes for the final result.

5.1 Security Server

The primary function of a Security server is to host one or more services for other hosts over a network. For instance, a file sharing services will be served by file server will delete, modify, access and store the files. A database server that serves database services for Web applications on Web servers. The Web servers provide Web related services to users Web browsers. There are many types of servers, such as authentication, directory services, application, email, logging, name/address resolution services, infrastructure management services such as Domain Name Server [DNS], print, and remote access.

This module provides background information on server security in the network. It depicts the process about categorizing the different needs of a security server, which results in identifying appropriate security controls on the network traffic. It discusses common server vulnerabilities, weakness and threats, by which servers will be placed in a secured environments.

5.2 File system Investigation module

In this module, Evidences will be extracted from the files which were deleted from the hard drive and the data reside on free spaces. This module depicts all the files and directories which are existing on hard drive. It will identify all the deleted files on the hard drive. From filing system, this tool gathers proof of deletion and search for data at deleted areas, free area, and slack areas. In the case of detection of free areas, this module identifies the main points of size used for data, size of free spaces and size of total hard drive.

5.2.1 Hidden Evidence Analysis in the File System Accused will try to hide the sensitive and confidential information at different locations of file system like bad clusters, Volume, file slacks and deleted free spaces.

The file system such as,

1) Hard Disk: In Hard disc the Protected Area on ATA disks/ maintenance track is used for hiding sensitive information. The evidence collection tools are used to copy these contents to do further investigation.
2) File System Tables: At NTFS file system, File Allocation Table Master File Table were utilized to keep track of files. Below Figure 2 depicts the MFT construction. MFT entries were utilized to hide important and sensitive data.

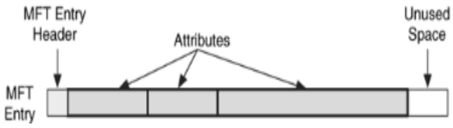


Figure 2: Master File Table (MFT) organization

3) File Deletion: In the operation of file deletion, the table record is taken out and it gives an impression to the user that, the file no longer exist. The file being deleted resided at the cluster as been marked as free so that, can be used to store new files. But, although the file is deleted, the data about the file is still available at the hard disk clusters. Information on files can be regain using the address of the start and end of the file in format of Hexadecimal and save by making a text file copy with corresponding extension.

Recovering a image file in the drive:

a) Opening a file which is in hexadecimal representation.

b) Investigate the corresponding signature of the file.

c) With start to end, signature of the file is copied.

4) Partition Tables: All the information about partition are kept in a partition table and it is a fragment of the Master Boot Record. When the system is booted, the partition table permits the system to know how the hard disk is prearranged and then passes this message to the os. Whenever a partition is removed, the partition table entry is also removed and the data is not possible to access. But, even though the partition entry is deleted, the deleted information still reside on the hard drive of the system.

5) Slack Space: The slack file is the space which is available at the end of the volume, these volume slacks are used to hide confidential information in the hard drive. The empty space between two partitions is also used to hide sensitive data. Figure 3 shows slack spaces in a Hard Disk. When any files are deleted in the hard drive, The file system will update all relevant information about deletion process, file details,date,time etc. The clusters that were previously allocated to storing are unallocated and can be reused to store a new information. But, the information which are present on the hard drive can recover immediately after it has been deleted. The deleted evidence still remain on the Hard disk until a new file overwrites them. However, if the new file does not take up the full cluster, a small portion of the old data might remain in the slack space. In this case, deleted file and partially overwritten file can be retrieved.

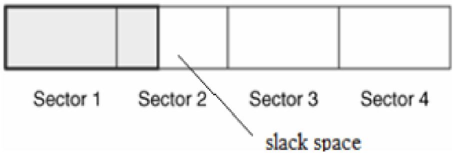


Figure 3: Slack spaces in a Hard Disk

5.3 Network Investigation

Network investigation module is equipped with a traffic monitoring tool for collecting important evidence in the network. A packet analyzer tool provides live forensic information about an attack. In order to monitor this process, system uses the Winpcap and Jpcap tools. These tools can extract Source IP, Destination IP , Source MAC, Destination MAC, Method, Protocol , Captured Time, Captured Length, Frame Type, Version and Destination Host. After completed this process save the file which contains all the forensic information.

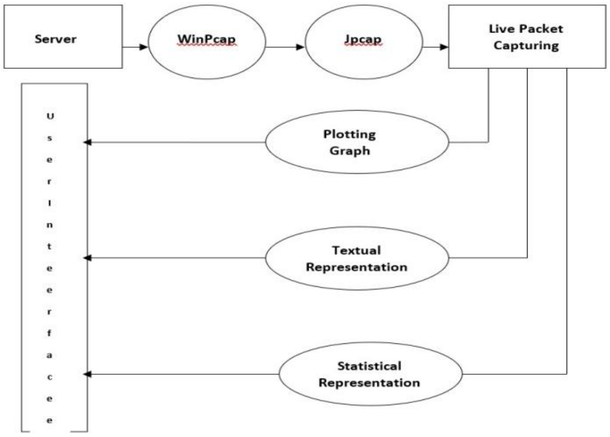


Figure 5: Network Forensics Architecture with each element

5.3.1 Server

In order to service requests, an application program called server which accepts connections to service requests by giving responses." A client application can also have server application running, or it may be connected via computer network. Database, VPN , DHCP , DNS , WINS , security servers, firewall, domain, proxy server, are examples

5.3.2 WinPcap

For packet capturing, Pcap is an application programming interface. In Windows, the pcap is a packet capturing tool called as WinPcap.In Windows,WinPcap is a open source tool to capture the live packets in the network. Through widely used system primitives, most networking applications access the network sockets. Easily transfer data on a network is achieved through this approach,

The need of WinPcap tool is to provide the access to Win32 applications, It provides following services to:

1. Raw live packets are captured, both from shared media and received media in the network.
2. On the network traffic collect statistical values like source address,destination address, Packet details etc.

By the way of a device driver, This set of information is collected, which is mounted at the networking portion of the kernels Win32 and a couple of DLLs. Through a powerful programming interface, all these features are exported and easily utilised by the applications.

5.3.3 Jpcap

Java class package called Jpcap which allows Java related applications to capture sending and receiving packets to the network. It supports Ethernet, ARP/RARP, TCP, UDP, IPv4, IPv6 and ICMPv4.

Jpcap tool will supports many packets: Ethernet, IPv4, ARP/RARP, IPv6, ICMPv4, TCP and UDP. Other different types of packets are captured as live raw packets, which contain the whole information of the packets.

5.3.4 Live Packet Capturing

By virtually, this digital investigation tool is designed to detect network traffic. Data scientist facilitate to monitor VPN (SSH / IPSec PPTP) and SSL (HTTPS) connectivity of wired/wireless networks. The graphical representation makes it easy to analyze information being sent between many computers.

Main functional features of the Winpcap and Jpcap tools are as follows:

1. Capturing Live Packets
2. Drawing Connection between hosts
3. Plotting Network Connection
4. Dumping Text information
5. Collecting Statistics of the live data
6. Determining IP Addresses

5.3.5 User Interface

This tool provides multiple windows to display data to the users. In order to implement user interfaces, Java's Swing class technology provides different classes. The user interaction is provides menu, where a user can select different windows, where each window represents information to the users.

These are:

- Statistical Information of information
- Textual representation of information
- Graphical representation of information

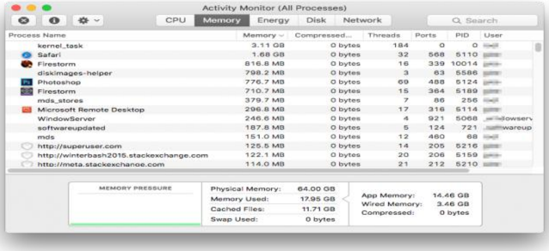
5.4 Memory Investigation

Memory Investigation process consist of the browser history and active process in the system. This process consists of lists all list all list out services running on system ,ports in the system, gives order in which system has booted up till that moment processes running on system, and login session details of the users. For implementing this we use the wmic tool.

5.4.1 PROPOSED FRAMEWORK FOR DATA COLLECTION PROCESS IN SYSTEM & RESULT

- Data Collection: In this process, the data related to source, destination addresses, time of activity, process ID, port address etc. will be collected for further processing.
- Graphical Interface: The graphical interface provides forensic examiners will assist in getting the information and proofs about the crime occurrence at memory. The results are displayed as per the user ridable form. The data collected can be presented in the form of tabular or pie chart.
- Relational Database: The proposed system will be used to store and process the needy and important information for investigation with the help of database. It is more flexible to handle the data, The information about safe user and the suspect will be identified by the attributes and user IP address.

Data at the Log: Log file data will be collected for processing, which contains the user behavior like page accessed, Ip address, Time of access and many more data, which will be collected and stored in the database for forensic.



Process Name	Memory	Compressed...	Threads	Ports	PID	User
kernel_task	3.11 GB	0 bytes	184	0	0	[redacted]
Safari	1.68 GB	0 bytes	32	568	5110	[redacted]
Firestorm	816.8 MB	0 bytes	16	339	10014	[redacted]
diskimages-helper	798.2 MB	0 bytes	3	63	5586	[redacted]
Photoshop	776.7 MB	0 bytes	69	488	5124	[redacted]
Firestorm	710.7 MB	0 bytes	15	364	5189	[redacted]
mds_stores	379.7 MB	0 bytes	7	86	256	[redacted]
Microsoft Remote Desktop	296.8 MB	0 bytes	17	316	5114	[redacted]
WindowServer	246.6 MB	0 bytes	4	921	5068	[redacted]dowserv
softwareupdated	187.8 MB	0 bytes	5	124	721	[redacted]wareup
mds	151.0 MB	0 bytes	12	460	68	[redacted]
http://superuser.com	125.5 MB	0 bytes	14	205	5216	[redacted]
http://winterbash2015.stackexchange.com	122.1 MB	0 bytes	20	206	5159	[redacted]
http://meta.stackexchange.com	114.0 MB	0 bytes	21	212	5210	[redacted]

MEMORY PRESSURE

Physical Memory: 64.00 GB
Memory Used: 17.95 GB
Cached Files: 11.71 GB
Swap Used: 0 bytes

App Memory: 14.46 GB
Wired Memory: 3.46 GB
Compressed: 0 bytes

Figure 6 : The running process in the computer system

The above figure shows, the running process in the system including running Threads, Ports, PID, usage of Memory in system and process name. The all process file related information is collected in text file for further investigation of the evidence.

5.5 Data Preprocessing

Data Pre-processing is the main which is process useful in generating models and patterns to from a enormous dataset. These models and patterns have an major role in a decision makings. Mining results are based purely on the Data quality. The data quality can be achieved by Processing the data to get required Information and removing unwanted information from the collection of the data. This is a major step to prepare raw data for getting the accurate results.

5.6 Mining Algorithm

1) Classify item sets/ variables and from a detailed case report. Our proposed system stores these variables as attributes of filesystem table, network table tables.
2) Consider the item sets I = {I1, I2, I3,I4 ⋯ Im}.
3) Set of actions D = {t1, t2, t3, t4⋯ tn}.
4) By using Apriori algorithm find frequent item sets. Find the set of frequent item sets. E.g. if an suspect attacked database, login attempt results a Data tampering /data loss and case report show actions like Login attempt , Data deleted, attack type = SQL injection. If frequent item sets are found then we can set a rule that, intention of suspect is stealing of data.
5) Rule of Association states that, if A →B association among A and B, such that, if A exist then Y will also exist. We can conclude that , if the accused handled files of OS, then ,intention behind this will be crashing the system.
6) Based on the Rule set ,formulate the SQL queries for database.
7) Data Recovery process.

VI. EXPECTED OUTCOME OF THE PROPOSEDWORK

- The data set can be obtained from any of the known source and system has the flexibility of importing data set from various sources.
- File system investigation module finding the evidence from the free spaces and deleted files.
- Network investigation in packet analyzer provides live forensic information about an attack.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

- Memory investigation process consist of the active process in the system and browser history of the suspect system.
- For crime data mining, system uses Apriori algorithms. This algorithm will predict the expected outcome by considering the outcome of File system analyzer, Network Forensic, Memory Forensic modules.

VII. CONCLUSION

- The threats in cyber transactions are increasing day by day at high rate. The data scientists are working on the cyber-crime prevention methodologies. Focus on digital forensic investigation is in very much need.
- This work addresses the issues of cybercrime suspect prediction by using data science methodologies, which adopts file system investigation, Network investigation, Memory Investigation and using Mining Algorithm to predict the cyber-crime suspect. This can be enhanced using high performance mining algorithms.

REFERENCES

1. Sindhu. K. K and Dr. B. B. Meshram,", A Digital Forensic Tool for Cyber Crimeining ",IRACST – Engineering Science and Technology: An International Journal (ESTIJ), ISSN: 2250-3498, Vol.2, No.1, 2012,pp.117-124.
2. Prashant K. Khobragade and Latesh G. Malik, "Data Generation and Analysis for Digital Forensic Application using Data Mining", 2014 Fourth International Conference on Communication Systems and Network Technologies,pp.458-462.
3. Sonal Honale and Jayshree Borkar, "Framework for Live Digital Forensics using Data Mining", International Journal of Computer Trends and Technology (IJCTT) – volume 22 Number 3–April 2015,pp.117-121.
4. Nikhil Kumar Singh,Deepak Singh Tomar and Bhola Nath Roy," An Approach to Understand the End User Behavior through Log Analysis", International Journal of Computer Applications (0975 – 8887) Volume 5– No.11, August 2010,pp.27-34.
5. V.Anitha and Dr.P.Isakki, "A Survey on Predicting User Behavior Based on Web Server Log Files in a Web Usage Mining", 978-1-4673-8437-7/16/$31.00 ©2016 IEEE.
6. Raburu George, Omollo Richard,Okumu Daniel," Applying Data Mining Principles in the Extraction of Digital Evidence", International Journal of Computer Science and Mobile Computing, Vol.7 Issue.3, March-2018, pp. 101-109.
7. Simon L. Garfinkel, "Digital Forensics Research: The next 10 years", Naval Postgraduate School, Monterey, USA, 2014. Pp.63-73.
8. **8.** Mandeep Kaur,Navreet Kaur,Suman Khurana,"A Literature Review on Cyber Forensic and its Analysis tools", International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 1, January 2016,pp.23-28.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication