2014 Fourth International Conference on Communication Systems and Network Technologies

Data Generation and Analysis for Digital Forensic Application using Data Mining

Prashant K. Khobragade
Dept. of Computer Sci. & Engi.
GHRCE, Nagpur-440016
prashulkhobragade@gmail.com

Latesh G. Malik
Dept. of Computer Sci. & Engi.
GHRCE, Nagpur-440016
latesh.malik@raisoni.net

Abstract- **In the cyber crime huge log data, transactional data occurs which tends to plenty of data for storage and analyze them. It is difficult for forensic investigators to play plenty of time to find out clue and analyze those data. In network forensic analysis involves network traces and detection of attacks. The trace involves an Intrusion Detection System and firewall logs, logs generated by network services and applications, packet captures by sniffers. In network lots of data is generated in every event of action, so it is difficult for forensic investigators to find out clue and analyzing those data. In network forensics is deals with analysis, monitoring, capturing, recording, and analysis of network traffic for detecting intrusions and investigating them. This paper focuses on data collection from the cyber system and web browser. The FTK 4.0 is discussing for memory forensic analysis and remote system forensic which is to be used as evidence for aiding investigation.**

Keywords: Data Collection; Log Data collection; Digital forensic tool; Clustering.

I. INTRODUCTION

Data mining technique have unlimited potential in the field of forensic science, where the models and tools can be developed to help investigators, digital forensics professionals and law enforcement officers to find the data or clues they are searching for much more efficiently and faster. As the technology increases the huge information is stored in digital form; in data mining the data generation, data warehousing and data analysis are the three important features involved in the examination process.

When crime occurs or involves the use of digital devices the investigation is categorized under digital forensic or cyber forensic. If the digital device involved is only a computer or digital storage medium, it refers to the investigation as computer forensic using only excel file in MSword, power point. It fails to uncover amount of evidence stored over time in the memory [9]. The various of digital devices used by individuals includes personal computers, wireless phones, cell phones, laptops, personal digital assistants (PDAs), wired landlines, broadband internet connection modems etc. Each person today maintains more than one mail account, which they have many communities, groups, takes active part in chat rooms and other social networking sites or blogs; with his/her identity or any other evidence about cyber crime in other digital storage media such as a flash drive [1][2]. The approach of the crowd sourced data in forensic investigation via the construction of a simple process model presented a simple model for crowd sourced digital forensics, and discussed various technique utilized in such forensic investigations. It generated data which supports the self information provides a good measure of the "uniqueness" of a given username, and that this "uniqueness" correlates with the amount of discoverable personal information [6].

Network forensics is used to find out attacker's behavior and tracked them by gathering and analyzing log information and status information of attacker's in network. The concept of Network Forensics occurs. Network forensic helps tries to analysis network traffic data, process in network on network devices like routers and switches. A forensics investigation requires the use of disciplined investigative techniques to discover and analyse traces of evidence left behind after a committed crime [7].

Network forensics is not providing the security on network or website. It is an extended phase of network security as the data for forensic analysis are collected from security products like firewalls and intrusion detection systems [3]. The results of this data analysis are utilized for investigating the attacks. However, there may be certain crimes which do not breach network security policies but may be legally prosecutable [3] [5].

Digital forensic contains computer forensics investigation, disk forensics analysis, network forensics analysis, firewall forensics, physical device forensics, database forensics, mobile device forensics analysis, software forensics examination, live systems forensics etc. Digital Forensic has been described as incident specific and practitioner driven advances which are developed and then applied [11]. When network forensic involves in cyber crime which deal with network traffic data monitored and try to find out malicious activity done on network or attacks made by hackers. If an attack is detected, then the nature of the attack is also determined. Network forensic techniques enable investigators to track back the attackers. The final goal is to provide evidence for the law enforcement [5] [10]. The concept of network forensics deals with data found across a network connection mostly access and routed traffic from one network to another. The data packet sent over the network has been captured by attackers. This may leads to intrusion detection in network forensic investigation [12].

978-1-4799-3070-8/14 $31.00 © 2014 IEEE
DOI 10.1109/CSNT.2014.97

In this paper network analysis involves collection of network traces and detection of attacks. The file system investigation is to indentified the evidence and collect it from storage device. The attackers has use various tool to penetrate website or network. The DOS attack may lead to crash down the site which is currently running on network, also SQlinjection attack has been very popular to database or website to alter the user credential. Our purpose is to fine out event occur in network and stored each event of user and attackers. The challenge is to identify useful network events and record minimum representative attributes for each event so that the least amount of information with highest probable evidence is stored. The critical step in the entire process of network forensics is to analyze attack data, arrive at a conclusion and finding the evidence.

II.LITERATURE SURVEY

The threat of having disruption due to cyber attacks has become a pressing issue. It has also become important to extract information from these huge databases that might be valued to the owner of the database. In network forensic analysis, it focused on the visual representation for network component for more easily analysis of evidence to the examiner. The visualization effect build to provide ease of understand the evidence or data [5]. Computer forensics is the process to collecting and analyzing evidence which is important and allowable for criminal investigation process. Data generation approach in physical storage device gives a unique way of generating data, storing data and analyzing data, which is retrieved from digital devices which pose as facts in forensic analysis. The approach used to recover such data from physical devices such as flash drives is recovered [2]. The common model proposed by Freiling and Schwittay in 2007, both for incident response and computer forensic processes, acceptable and manageable approach in the digital forensic investigations, while retain the opportunity of a exact forensic investigation[13].

Network forensics among them is used to find out attackers behavior and tracked them by collecting and analyzing log data and status information of attacker' [10]. A general approach to the forensic research is to find specific text strings by comparing every byte of the digital evidence at the physical level [8]. This paper focuses on investigation of data capture from the computer system. The logical memory data is capture and also the activity of running process can able to analyze. The data generation and analysis of file captured in computer system. In network traffic forensic the log files analysis and file system forensic analysis take place using forensic application.

The collection of data is significant for every crime investigation process, when there are so many attacks happened in network or any website. During analysis of log data and find out malicious behavior activities of user or attacker's data by mining log file from database. Access logs can contain plenty amount of data regarding each user activities [11].

III. PROPOSED FRAMEWORK FOR DATA COLLECTION PROCESS IN SYSTEM

A. Data Collection:

In the Data collection step the forensic process is to identify possible information of source node and collect all information related to source, destination, and time of activity, process ID, port address. Major sources of data are personal computers, browsing log data information, digital storage media which store image of data, Routers in network device, Cell Phones, Digital Camera, Network traffic in system etc. [4]

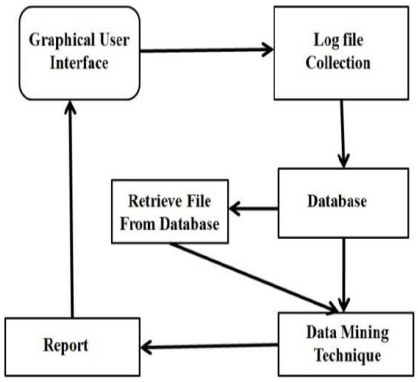


Figure 1. Block Diagram of proposed system

The proposed system is planning to develop and obtains data according to their importance, unpredictability and amount of effort to collect data.

B.Graphical Interface:

The graphical interface provides investigators to provide ease in find out the evidence from the computer system where crime is occur. This is also shows the result of processing data in presentable form. The representation of data can be able to analysis with using pie chat, tabular form or in form of report.

C.Database:

Database is used to store and collects evidence from crime side using proposed system and stored required data for investigation. The database provides flexibility to store the data in table format. The attribute gives more information about the normal user and the malicious user, also it able to know the source IP address of user.

D.Log File Collection:

In this step log data of browser have been captured and stored in database for investigation. This file log contains history of browser in which user log information retrieved.

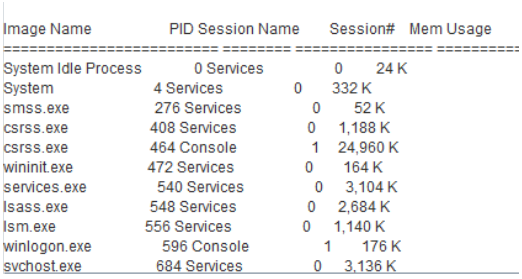
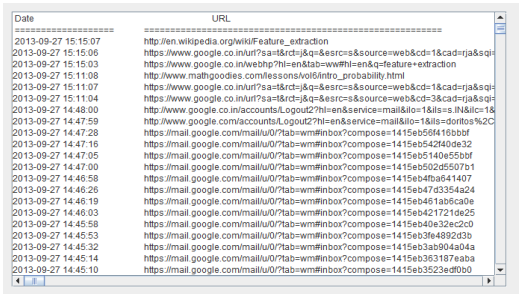


Image Name	PID	Session Name	Session#	Mem Usage
System Idle Process	0	Services	0	24 K
System	4	Services	0	332 K
smss.exe	276	Services	0	52 K
csrss.exe	408	Services	0	1,188 K
csrss.exe	464	Console	1	24,960 K
wininit.exe	472	Services	0	164 K
services.exe	540	Services	0	3,104 K
lsass.exe	548	Services	0	2,684 K
lsm.exe	556	Services	0	1,140 K
winlogon.exe	596	Console	1	176 K
svchost.exe	684	Services	0	3,136 K

Figure 2- Running process in system

The figure 2 shows the running process in the computer system including running process name, session name and usage of memory in system. The all process file related information is retrieve in text file for further examination analysis of evidence.



Date	URL
2013-09-27 15:15:07	http://en.wikipedia.org/wiki/Feature_extraction
2013-09-27 15:15:06	https://www.google.co.in/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&sqi=
2013-09-27 15:15:03	https://www.google.co.in/webhp?hl=en&tab=ww#hl=en&q=feature+extraction
2013-09-27 15:11:08	http://www.mathgoodies.com/lessons/vol6/intro_probability.html
2013-09-27 15:11:07	https://www.google.co.in/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&sqi=
2013-09-27 15:11:04	https://www.google.co.in/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&cad=rja&sqi=
2013-09-27 14:48:00	http://www.google.co.in/accounts/Logout2?hl=en&service=mail&ilo=1&ils=s.IN&ilc=1&
2013-09-27 14:47:59	http://www.google.com/accounts/Logout2?hl=en&service=mail&ilo=1&ils=doritos%2C
2013-09-27 14:47:28	https://mail.google.com/mail/u/0/?tab=wm#inbox?compose=1415eb56f416bbbf
2013-09-27 14:47:16	https://mail.google.com/mail/u/0/?tab=wm#inbox?compose=1415eb542f40de32
2013-09-27 14:47:05	https://mail.google.com/mail/u/0/?tab=wm#inbox?compose=1415eb5140e55bbf
2013-09-27 14:47:00	https://mail.google.com/mail/u/0/?tab=wm#inbox?compose=1415eb502d5507b1
2013-09-27 14:46:58	https://mail.google.com/mail/u/0/?tab=wm#inbox?compose=1415eb4fba641407
2013-09-27 14:46:26	https://mail.google.com/mail/u/0/?tab=wm#inbox?compose=1415eb47d3354a24
2013-09-27 14:46:19	https://mail.google.com/mail/u/0/?tab=wm#inbox?compose=1415eb461ab6ca0e
2013-09-27 14:46:03	https://mail.google.com/mail/u/0/?tab=wm#inbox?compose=1415eb421721de25
2013-09-27 14:45:58	https://mail.google.com/mail/u/0/?tab=wm#inbox?compose=1415eb40e32ec2c0
2013-09-27 14:45:53	https://mail.google.com/mail/u/0/?tab=wm#inbox?compose=1415eb3fe4892d3b
2013-09-27 14:45:32	https://mail.google.com/mail/u/0/?tab=wm#inbox?compose=1415eb3ab904a04a
2013-09-27 14:45:14	https://mail.google.com/mail/u/0/?tab=wm#inbox?compose=1415eb3631d7eaba
2013-09-27 14:45:10	https://mail.google.com/mail/u/0/?tab=wm#inbox?compose=1415eb3523edf0b0

Figure 3- Log file data from web browser

Figure 3 shows log information data from the web browser which include browsing date, URL that are recently browsed. In crime investigation log history of web browser gives more information about thread or normal user file analysis.

E.Data Mining:

The useful clustering technique has used to form a similar kind of group. With clustering technique it can be possible to categorized normal user data and the attack in network.

IV- GENERAL MEHODOLOGY FOR DIGITAL FORENSIC EXAMINATION

A. Preparation

In forensic examination data preparation is important for investigator; this data can be obtained as user log data file and tracing the attackers or authenticated user or attackers. The data is obtained and stored for backup in storage devices. A hash of all the trace data is preserved. A copy of the data will be analyzed and the original network traffic data which is not alter by hacker.

B. Detection

The presence and nature of the attack are determined by the parameter like IP address, port, process ID, session ID. A quick validation is done to assess and confirm the suspected attack.

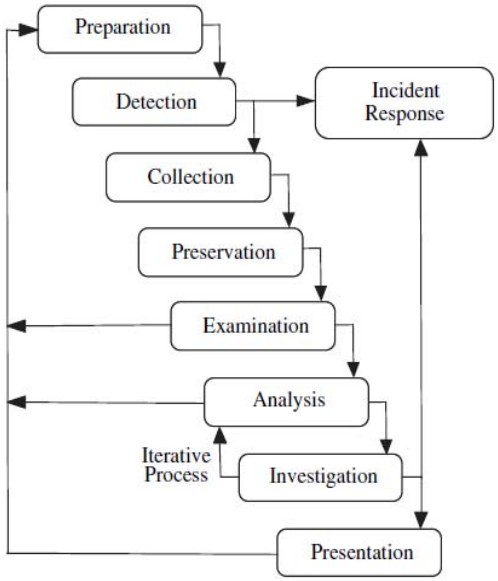


Figure 4. General forensic model

C. Generation

Data are acquired and used to collect the traffic data. This leads to generation of hugh amount of data which requires amount of storage memory and the system must be able to handle different log data formats appropriately.

D. Examine

The traces obtained from various nodes which are integrated and combine to form one large data on which analysis is performed. There will be some issues like redundant information and overlapping time zones which need appropriation.

E. Analysis

The indicators are classified and correlated to deduce important observations using the existing attack patterns. The attacks patterns are put together reconstructed and try to understand what the intention of attacker and methodology used is.

F. Investigation:

The goal is to determine the path from a victim network or system through any intermediary systems and communication pathways, back to the point of attack beginning. The packet captures and statistics obtained are used for attribution of the attack.

G. Presentation:

The remark are be presented in a user friendly format to understandable for investigator, while providing explanation of the various procedures and information used to arrive at the conclusion. The conclusions are also presented using presentable format so that they can be easily grasped and make their decision.

V. DATA MINING TECHNIQUE

Data mining technique is essential to a data mining system and ideally consists of a set of functional modules for tasks such as characterization, association, cluster analysis, classification, prediction, regression. The clustering technique gives similar kind group at which a group of cluster is formed as normal user and the attacker's.

A. Clustering

Data clustering, builds unsupervised data models. Data instances are grouped together, based on similar kind matching schemes, defined by the clustering system, in large, multi-dimensional data set. As clustering attempts to group data instances into clusters of significant interest, assess the performance of the model and detect attackers.

In this approach clustering has been used as a step in analysis of the data generated as part of the digital forensic examination. Our significance is to examine those data instances which do not group naturally into cluster to form groups, for forensic evidence. The used of simple k-means algorithm for the basic clustering on generated data. Simple k-means algorithm takes k, the number of clusters in the unorganized data to be calculated, as an input parameter and partitions the given set of n objects into k clusters so that the resulting intra cluster similarity is high while the inter cluster similarity is low. Euclidean distance measure is used to assign instances to clusters.

VI. DIGITAL FORENSIC TOOLKIT FTK 4.0

Forensic Tool Kit is a commercial forensics tool developed by Access Data. FTK is accepted for digital investigations in various domain of cyber crime, using the FTK platform built provide speed, stability and ease of use as there other product in the market. FTK has database driven, enterprise-class architecture allows handling massive data sets, as FTK provides stability and processing speeds which not possible with other tools.

A. Features of FTK:

- Include integrated computer forensics solution.
- Create images, process a wide range of data types from forensic images to email archives, analyze the registry, conduct an investigation
- FTK Architecture provides better Stability to the examiners.
- FTK is database driven.
- Broad file system, file type and email support
- Support for 700+ image, archive and file types.
- Data Visualization
- Advanced volatile / memory analysis
- The Static RAM memory analysis from an image or against a live system analysis.

B.Use of FTK4.0 in Remote Investigation

FTK4.0 provides remote investigation for single node examination and visualization analysis. In remote data investigation includes to analyze process information, service information, driver information, network device, network information. In network forensic, agent is push to remote machine by entering IP address of remote machine and with valid credential of that remote machine.

The examination take place with connection of remote machine and evidence collection has been start for which include volatile data, memory data and drive data. The socket gives information about connectivity of remote network, once it possible to determine connectivity to network then it easily find whether it is thread or not. It also possible to see the remote system service list, user account, administrator account information.

C. Memory Forensic:

The building method for memory acquisition and memory analysis is very crucial task and memory analysis is done quickly and more accurately when there are security breaches by attacker's or unauthorized person, all necessary information is much need for investigation purpose. FTK provides acquire physical memory, volatile memory analysis. With volatile memory it shows the process which is run on the system, also shows the hidden files in the physical memory. Windows is not aware about the labeled as suspicious and highlighted data or files in system, so the investigator able to see what processes are hidden.

Memory analysis is performed on single system or the agent system where the investigation took placed. If the memory analysis is performed on agent then agent has to hold the string given by the examiner to find out related items which has to analysis and later it return to examiner for further analysis of this data. During analysis examiner can investigate respective related items which is captured from agent system, and if any event found to real or crime in those data or hidden files. The examiner has also been find out all individual files or directories form the physical memory for further analysis, once the file is captured from the memory then examiner investigate all relating data for files and made work flow for ease of use.

When working analysis is done with file in memory is easy to generate statistical report form examiner system, this report provides all working files, working directory, document, evidence report. This report evidence provides leads to the law enforcement.

CONCLUSION

In this paper cyber crime data is collected with using of proposed methodology, the log file is captured and stored in database as evidence. The generated file and data is analysis with using digital forensic toolkit. Forensic toolkit is used t analysis of victim system where the attack is happened. The physical memory data and logical memory data is analyze and find an evidence which help in crime investigation.

In future work, forensic toolkit is used to analyze remote node and easily to visualize data for report presentation.

REFERENCES

[1] Jooyoung Lee, Sungkyung Un, and Dowon Hong, "Improving Performance in Digital Forensics", *International Conference on Availability, Reliability and Security,* 2009.

[2] Veena H Bhat, Prasanth G Rao, Abhilash R V,P Deepa Shenoy,Venugopal K. R. "A Novel Data Generation Approach for Digital Forensic Application in Data Mining", *Second International Conference on Machine Learning and Computing*, 2010.

[3] Sebastian Schmerl, Michael Vogel, René Rietz, and Hartmut König, "Explorative Visualization of Log Data to support Forensic Analysis and Signature Development", *Fifth International Workshop on Systematic Approaches to Digital Forensic Engineering*, 2010.

[4] Funminiyi Olajide, Nick Savage, Richard Trafford, "Forensic Memory Evidence of Windows Application", The 7th International Conference for Internet Technology and Secured Transactions (ICITST-2012).

[5] Seung-hoon Kang, Juho Kim, "Network Forensic Analysis Using Visualization Effect", International Conference on Convergence and Hybrid Information Technology, 2008.

[6] Daniel Compton, J.A. Hamilton, "An Examination of the Techniques and Implications of the Crowd-sourced Collection of Forensic Data", *IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing*, 2011.

[7] Meixing Le, Angelos Stavrou, Brent Byunghoon Kang, "Doubleguard: Detecting Intrusions In Multitier Web Ap p lications", *IEEE transactions on dep endable AND secure computing, vol. 9, no. 4, pp 512-525, july/august 2012.*

[8] Jooyoung Lee, Sungkyung Un, and Dowon Hong, "Improving Performance in Digital Forensics", *International Conference on Availability, Reliability and Security,* 2013.

[9] Lianfi Y in, "Research on windows physical memory forensic analysis", *Fourth International Symposium on Information Science and Engineering,* 2012.

[10] X. YIN, W. Yurick, M. Treaster, Y. Li, and K. Lakkaraju, "VisFlowConnect: NetFlow Visualizations of Link Relationships for Security Situational Awareness", *ACM Workshop on Visualization and Data Mining for Computer Security (VizSec), Washington, D.C.*, October 2004

[11] Mohd Taufik Abdullah, Ramlan Mahmod, Abdul A. A. Ghani, Mohd A Zain and Abu Bakar Md S, "Advances in Computer Forensics," *International Journal Of Computer Science and Network Security*, vol. 8, no. 2, February 2008.

[12] Kara Nance, Brian Hay and Matt Bishop, "Digital Forensics: Defining a Research Agenda," *Proc. of the Forty Second Hawaii International Conference on System Sciences*, pp. 1-6, 2009.

[13] F. C. Freiling, and B. Schwittay, "A Common Process Model for Incident Response and Computer Forensics," *Proc. of Conference on IT Incident Management and IT Forensics, Germany,* 2007.