A Review on Data Generation for Digital Forensic Investigation using Data Mining

[1] Prashant K. Khobragade, [2] Latesh G. Malik

[1] PG Student, Department of CSE, GHRCE,
Nagpur-440016, Maharashtra, India

[2] Professor, Department of CSE, GHRCE,
Nagpur-440016, Maharashtra, India

Abstract - Digital forensic is part of forensic science that unconditionally covers cyber crimes. In a cyber crime digital forensic evidence examination requires a special process and techniques in examination of cyber crime in crime scene and examination of evidence are accepted in law enforcement. Cyber crime involves log data, transactional data is occurs which tends to plenty of data for storage and analyze them. The network forensic traces involve 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 this paper general methodology is discussed for network data forensic analysis and also the survey of various network forensic analysis tools and approach in use to capturing data from different resources.

***Keywords* - Data Collection, Network Forensic Tools, Digital Forensic.**

1. 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; data Preparation/ Generation, Data warehousing and Data Mining, are the three essential features involved in the investigation process. When crime is aided by 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, we refer to the investigation as computer forensic. Computer forensic (aka digital forensic) is a branch of forensic science, whose goal is to explain the current state of the digital artifact [2]. The pool of digital devices used by individuals includes cell phones, laptops, personal digital assistants (PDAs), personal computers, wireless phones, wired landlines, broadband/satellite internet connection modems, iPods etc. Each individual today maintains more than one email account, is a member of many communities, virtual groups, takes active part in chat rooms and other networking sites with his/her identity or under an alias, juggles multiple flash drives and other digital storage media [1] [2]. The concept of the crowd sourced 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 [6].

Network forensics is not another term for network security. 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. 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. The concept of network forensics deals with data found across a network connection mostly ingress and egress traffic from one host to another. Network forensics tries to analyze traffic data logged through firewalls or intrusion detection systems or at network devices like routers and switches [5].

Network forensics involves network traffic monitoring and determining if there is an anomaly or intruder in the traffic and ascertaining whether it indicates an attack in cyber space. If there is an attack is detected, then the nature of the attack is also determined with specialized technique. Network forensic techniques enable investigators to track back the attackers. The resulting goal is to provide sufficient evidence to allow the perpetrator to be prosecuted in law of court [5][10]. The clustering algorithm is typically used for exploratory data analysis, where there is little or no prior knowledge about the data [11].

Much of the data in those files consists of unstructured text/data whose data analysis by examiner, which is difficult to be performed and consume lots of time to get clue for further investigation. The clustering algorithm is the data within valid cluster having more similar characteristic of information [11].

2. Literature Survey

There are few studies relating to data generation, such as data collection from hard drive, flash drive, and crowed source data. The forensic analysis is also made with some of the data mining tools as WEKA for visualization of data. In network forensic analysis, it emphasis on the visualization effect related to network components is made to build more efficient forensic system [5]. Computer forensics is the process that applies computer science and technology to collect and analyze evidence which is important and admissible to criminal investigations. Data generation [2] 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 common model proposed by Freiling and Schwittay in 2007, both for incident response and computer forensic processes, permitted a management oriented approach in digital investigations, while retain the opportunity of a exact forensic investigation [10].

Network forensics among them is used to find out attackers behavior and trace them by collecting and analyzing log and status information [10]. A [8] general approach to the forensic research is to find specific text strings by comparing every byte of the digital evidence at the physical level for computer forensic investigation.

3. Survey on Data Collection from Different Resources

Digital forensic is part of forensic science that implicitly covers crime that is related to computer technology. In a cyber crime, digital evidence investigation requires a special procedures and techniques in order to be used and be accepted in court of law.

The data capturing from physical memory as hard drive is to be collect data that not yet been overwrite on same memory location of that hard drive. There are several software tools in a market to retrieval of data [9]. This data collection mainly focuses on windows like system using method based on hardware and software for system. Data gathering on such physical drive having some issues as lack of reliability, more consume time to retrieval of data, the software tools that unavoidably destroy the overwrite data on physical memory.

The WEKA tool is used to analyze data from flash drive, the tool is open source or those already existing in the present day computing environment, hence easily accessible to the investigation [2]. A statistical approach is used in validating the reliability of the preprocessed data. It also forms a bridge between the digital forensic investigation team and judicial bodies [2].

4. Survey of Network Forensic Analysis Tools (NFATS)

4.1 TCPFlow

Captures data transmitted as part of TCP connections (flows) and stores it for protocol analysis. It reconstructs actual data streams and stores in a separate file. TCPFlow understands sequence numbers and will correctly reconstruct data streams regardless of retransmissions or out-of-order delivery.

4.2 Flow-Tools

Library to collect, send process and generate reports from Net Flow data. Important tools in the suite are flow capture which collects and stores exported flows from a router, flow-cat concatenates flow files, flow report generates reports for Net Flow data sets, and flow-filter filters flows based on export fields.

4.3 TCPReplay

TCPReplty is a suite of tools with ability to classify previously captured traffic as client or server, rewrite layer 2, 3 and 4 headers and finally replay the traffic back onto the network.

4.4 Snort

Network intrusion detection or prevention system capable of performing packet logging, sniffing and real-time traffic analysis. It can perform protocol analysis, content searching, matching and application-level analysis.

4.5 Ntop

Used for traffic measurement, network traffic monitoring, optimization, planning, and detection of network security violations. It provides support for both tracking ongoing attacks and identifying potential security holes including IP spoofing, network cards in promiscuous mode, denial of service attacks, Trojan horses and port scan attacks.

5. Methodology for Network Forensic Examination

5.1 Preparation

The original data obtained in the form of traces and logs are stored on a backup device like read only media. 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.

5.2 Detection

The presence and nature of the attack are determined from various parameters. A quick validation is done to assess and confirm the suspected attack.

5.3 Generation

Data are acquired and used to collect the traffic data. The amount of data logged will be enormous requiring huge memory space and the system must be able to handle different log data formats appropriately.

5.4 Examine

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

5.5 Analysis

The indicators are classified and correlated to deduce important observations using the existing attack patterns. The attack patterns are put together, reconstructed and replayed to understand the intention and methodology of the attacker.

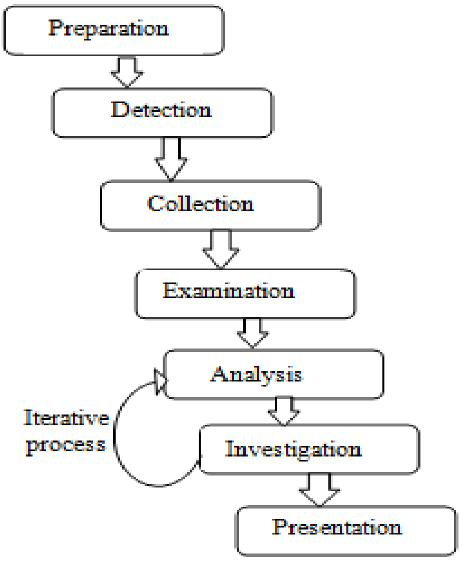

Figure 1.General Network forensic model

5.6 Investigation

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

5.7 Presentation

The observations are presented in an understandable language for legal personnel while providing explanation of the various procedures used to arrive at the conclusion. The conclusions are also presented using visualization so that they can be easily grasped.

6. Research Challenges

The frameworks for network forensic analysis tools have been discuss in the previous section. The limitations and specific research gaps associated with different phases in each implementation are given as:

6.1 Data Collection

The data collection from different resources involves challenge and identifies useful network procedures and record minimum representative attributes for each incident action so that the smallest amount of information with highest probable evidence is stored.

6.2 Data Examination

The data collected from various tools or resources must be aggregated and examined to discover whether investigation should be commenced. Data fusion of all the logs collected from various security tools deployed in each hosts on the entire network is a crucial problem.

6.3 Data Analysis

The data analysis is important step in the entire process of network forensics is to analyze attack data and arrive at a conclusion, pointing at the source. Classification and clustering of network events need to be done so that scrutiny of large volumes of data to understand their relationship with attacks becomes easy.

6.4 Instance Response

The real-time response to the network misuse is to be performed so that important data are not lost by the time response is initiated. The response processes are to be launched immediately when alerts begin.

7. Conclusion

This paper briefly describes various resources where the data is captured and also the ideas behind different tool in network forensic. Without getting into too many technical details, it modestly glances over the ideas and notions of some tools, and also explains framework for Network forensic.

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] Kyriacos E. Pavlou and Richard T. Snodgrass, Senior Member, IEEE, "The Tiled Bitmap Forensic Analysis Algorithm", *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,* VOL. 22, NO. 4, APRIL 2010.

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

[9] 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.

[10] 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.

[10] Nikkel BJ "Generalizing sources of live network evidence. Digital Investigation" , *International Journal of Digital Forensics & Incident Response,* vol. (3):193–200, Sept. 2005.

[11] Luís Filipe da Cruz Nassif and Eduardo Raul Hruschka, "Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection", *IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY,* VOL. 8, NO. 1, JANUARY 2013.