A Data Fusion based Digital Investigation Model as an Effective Forensic Tool in the Risk Assessment and Management of Cyber Security Systems

Ms. Suneeta Satpathy (PhD Scholar, Utkal University)
Assistant Professor, College of Engineering
Plot-1, Sector-B, CNI Complex, Patia, Bhubaneswar-24, Orissa, India
and
Mr. Asish Mohapatra, MSc, MPhil (pre-doctoral), EMC, Risk Cert (Harvard)
Regional Health Risk Assessment and Toxicology Specialist, Health Canada (Alberta region)
Suite 282, 220-4th Ave SE, Calgary, Alberta, Canada, T2G 4X3

ABSTRACT

The cyber-infrastructure have become increasingly complex and inextricably intertwined with the infrastructures of the public, and private organizations. The inter-connectedness of computers globally has enhanced our capability to analyze various databases; however, it has also raised the issue of information and databases security on the web. The law enforcement and the computational forensic analysis process, in its relative infancy, is the unwilling victim of the rapid advancement of information technology. An epistemic uncertainty is an unavoidable attribute which can be present in digital investigations and could affect the investigation process. So there has to be a well-designed system to analyze information from various sources for possible security threat and act appropriately in the event of any suspicion. Forensic digital analysis is unique among all the forensic applications. It is inherently mathematical and generally comprises of more data from an investigation than other types of applications. In this paper, we have presented a data fusion based digital investigation model by which conflicting information due to the unavoidable uncertainty can be identified and processed. Data fusion along with data mining techniques applied in the context of database and intelligence analysis can be correlated with various security issues and crimes. Thus it holds the promise of alleviating such problems. Application of our proposed model in the broader Health Care and Life Sciences (HCLS) and public health risk analysis (PHRA) and toxicological database integration areas are briefly discussed and future projects are proposed.

Keywords: Information Technology, Data Mining, Data Fusion, Cyber Crime, Digital Evidence, Computer Forensics (CF).

1. INTRODUCTION

Undoubtedly, the World Wide Web (WWW) connectivity through the ever-growing cyber-infrastructure has facilitated rapid availability of information resources and databases. As a result, there is a tendency among the users, business enterprises, private and public sectors to get connected to the WWW in order to take advantage of both the pull and push technology. There is a tremendous growth in computer and web related crimes; while, a similar growth is missing in the development of security solutions. It has created a new type of warfare where information systems, web based databases are the targets by which it is being exploited by the unscrupulous elements in the society for disrupting peace and causing mayhem **[5].** The criminal justice delivery system has not kept pace with the technological advancements, which have taken place with the advent of Information technology. To effectively combat cyber-infrastructure related crimes, it is not only sufficient to successfully investigate the crime but more important is to prosecute and administer justice, according to the law of the land.

Based on current approaches, security systems and databases generate enormous amounts of data and therefore, higher priority must be given to systems that can analyze rather than merely collect such data, while still retaining collections of essential forensic data. The forensic aspect to the overall model of security is equally important as the area of Computer Forensics (CF) lends itself heavily to the response of a criminal violation that has already occurred. Data fusion promises to play a proactive and central role in the future prevention, detection, attribution, and remediation of such types of cyber crimes.

Our research is aimed at providing a fusion based digital investigation model which can address various types of threats in network and facilitate forensic analysis. The model called "**A fusion based investigation model for Computer Forensic**s"[**3**]**,** is derived from the Joint Director Laboratories (**JDL**) data fusion model [**7**] and is built around a set of algorithms in various levels of fusion. The algorithms at various levels can be executed continuously and autonomously in its environment, able to carry out activities in a flexible and intelligent manner while being responsive to changes in its environment.

This paper concerns CF as it is a problem of great significance to information infrastructure protection because computer networks are at the core of the operational control of much of the day to day operations.

2. COMPUTER FORENSICS (CF) [1,2]

CF is the science of busting cyber criminals. It can be defined more pedantically as the "investigation of digital evidence for use in criminal or civil courts of law." CF is most commonly used after a suspected hack attempt, in order to analyze a computer or network for evidence of intrusion. It is the use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation, and for the purpose of presentation of digital evidence derived from digital sources in the court of Law to punish the criminal [**1**]. The major goals are to:

- Provide a conclusive description of all cyber-attack activities for the purpose of complete post-attack enterprise and critical infrastructure information restoration;
- Correlate, interpret, and predict adversarial actions and their impact;
- Make digital data suitable and persuasive for introduction into a criminal investigative process; and
- Provide sufficient evidence to allow the criminal perpetrator to be successfully prosecuted.

A major issue to achieve these goals is how to rapidly collect and normalize digital evidence from a variety of sources including firewalls, hosts, network management systems, and routers. The information that is collected could then be used to predict or anticipate adversarial actions, understand the current state of affairs, and help in determining appropriate courses-of-action.

2.1 Problem Statement

CF faces several problems. Some of them are highlighted below.

- Digital investigations are becoming more time consuming and complex as the volumes of data requiring analysis continue to grow.
- Digital investigators are finding it increasingly difficult to use current tools to locate vital evidence within the massive volumes of data.
- Log files are often large in size and multi-dimensional, which makes the digital investigation and search for supporting evidence more complex.
- Digital evidence [**6, 8**] by definition is information of probative value stored or transmitted in digital form. It is fragile in nature and can easily be altered or destroyed. It is unique when compared to other forms of documentary evidence.
- Computer forensic tools available are unable to analyze all the data found on computer system to reveal the overall pattern of the data set, which can help digital investigators decide what steps to take next in their search. Also the data offered by computer forensic tools can often be misleading due to the dimensionality, complexity and amount of the data presented.

Our proposed data fusion based investigative tool will concentrate on improving the quality of data rather than quantity for analysis. Further, it will lessen the processing time required and ultimately reduce the monetary costs of digital investigations.

3. DATA FUSION SYSTEMS AND RELATED WORK

Multi-sensor data fusion is an evolving technology, concerning the problem of how to fuse data from multiple sensors in order to make a more accurate estimation of the environment and to generate information of a superior quality [**7**, **11, 12**]. It is a formal framework in which the means and tools for the alliance of data originating from different sources are expressed [**14**]. The first data fusion methods were primarily applied in the military domain, in recent years these methods have also been applied to problems in the civilian domain and various non-military applications (e.g., air traffic controls, robotics, image processing, remote sensing, hazardous wastes tracking, environmental data fusion, etc.). A more recent idea is the application of Multisensor data fusion techniques to the area of information security [**13, 16**].

Multi-sensor data fusion provides an important functional framework for building next generation security systems. Tim Bass presented a Data Fusion model, based on the Joint Directors of Laboratories (JDL) Functional Data Fusion Process Model [**16**].

There are a number of research projects that have started to implement Multisensor data fusion techniques. One of these projects is EMERALD [**4**], an acronym for `Event Monitoring Enabling Responses to Anomalous Live Disturbances'. It couples sensors, so the state of one sensor can adjust another. This suppresses false positives and increases sensitivity. The idea in all types of fusion seems to use an approach, which mainly focuses on the implementation. There is no general architecture of Forensic data fusion systems. Based on these observations, it seems important to start a systematic analysis and to develop a generic architecture for forensic fusion-based investigation model, which can facilitate digital forensic analysis and to ultimately restrain the cyber criminals.

3.1 Requirements of Forensic Data Fusion system

Developing a system for agencies conducting CF investigations that will utilize the data fusion technology requires an appropriate methodology for selecting architecture and adopting alternative techniques for cost-effective system requirements [**7**]. Generally accepted engineering guidelines for data fusion systems recommend a paradigm in which the design and development flow from an overall system requirements and constraints to a specification of the role for data fusion within the system. There are several fundamental issues, which should be taken into consideration when building an investigation model [**9, 10**]:

- What architecture should be used?
- What algorithms and techniques are appropriate and optimal for a particular application?
- How should the individual source data be processed to extract the maximum amount of information?
- How does the data collection environment affect the processing?
- How can the fusion process be optimized?
- What accuracy can be achieved by a data fusion process?

Contemporary view on the problem of security is concerned with an idea that particular protective mechanisms and corresponding software must be integrated along with the forensic capabilities into a fusion system interacting via exchange of information and making decisions in a cooperative and coordinated manner. These systems should be adaptive to traffic variations, reconfiguration of the software and hardware components.

4. A FUSION BASED DIGITAL INVESTIGATION MODEL

As mentioned above, the proposed model (figure 1), "**A fusion based Investigation Model for Computer Forensics"[3]** is motivated by Data fusion model proposed by the JDL **[**7] that fuses data from various heterogeneous sources in order to attain low false alarm rates and high threat detection rates. In addition to the functionalities of Data Fusion Model, our model also supports post mortem forensic analysis by preserving the necessary potential legal digital evidence. Therefore, the main goal is to provide proactively a more intelligent fusion based model for CF which partly automates the detection and prevention of intrusions and handles the true positives there by reducing the number of events the operator of the system has to inspect as well as prioritizing these events. The data fusion process is further explained in four different progressions.

- Collection of events from different sources
- Processing of events in various levels of fusion
- Decision making
- Evidence accumulation

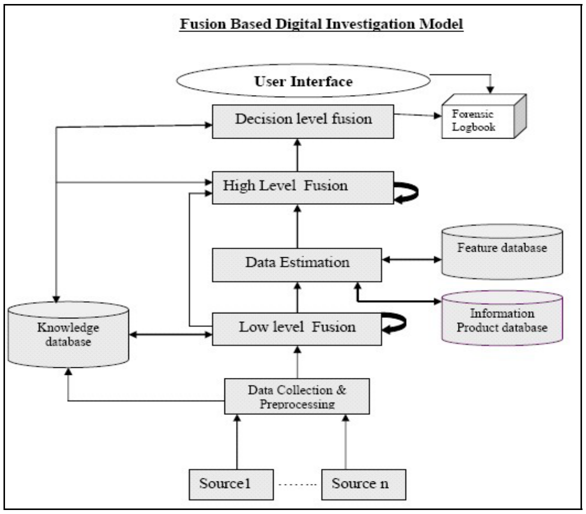

Figure 1 – A fusion based digital investigation model

4.1 Data Collection & Pre-Processing

The first step is the data collection & preprocessing phase where data collected from various sources are fused and processed to produce data specifying semantically understandable and interpretable attributes of objects [**7**]. The collected data are aligned in time, space or measurement units and the extracted information during processing phase is saved to the knowledge database or knowledgebase.

4.2 Low level fusion

This level of fusion processes the data to achieve a refined representation and reduces the quantity by concurrently retaining useful information and improves its quality, with minimal loss of detail. It is mainly concerned with data cleaning, data transformation and data reduction [**7**].

- Data cleaning (removes irrelevant information)
- Data transformation (converts the raw data into structured information)
- Data reduction (reduces the representation of the dataset into a smaller volume to make analysis more practical and feasible)

The above procedures enable the data fusion process to focus on data that applies most to the current situation and reduces the data fusion system load. It can help reduce a search space into smaller, more easily managed parts which can save valuable time during digital investigation.

4.3 Data estimation phase

It is based on a model of the system behavior stored in the feature database and the knowledge acquired by the knowledgebase. The fusion algorithm estimates the state of the system. After extracting features from the structured datasets, data fusion system will save them to an information product database.

4.4 High-level fusion

This level of fusion develops a background description of relations between entities. It consists of event and activity interpretation and eventually contextual interpretation [**7**]. Furthermore, it involves the use of data mining functionalities such as classification and clustering to extract useful patterns among the data. The results obtained would be indicative of destructive behavior patterns. These features form a feature space fused to identify and classify them to serve for attack detection and recognition. It effectively extends and enhances the completeness, consistency, and level of abstraction of the situation description.

4.5 Decision level fusion

The patterns discovered from the high level fusion still needs to be analyzed to determine the relevancy of those patterns. The goal of this step is to identify pertinent patterns. Further, it analyzes the current situation and projects it into the future to draw inferences about possible outcomes. It identifies intent, lethality, and opportunity [**7**]. Finally, decision of the fusion result along with necessary information is stored in the forensic logbook from which the forensic evidence report can be generated to be used in the expert testimony in the Court of Law.

4.6 Forensic Logbook

The forensic log book is a record keeping system. The term forensics refers to the post-mortem analysis of evidence; however, in the context of computer we refer to the analysis of evidence as CF [**1, 2**]. The digital information captured are recorded in the log book with a pre-defined format like date and time of the event, intruder's IP address, and target IP address, users, type of event, and success or failure of the event, origin of request for identification/authentication data and name of object for object introduction and deletion data. A time stamp is added to all data logged. The time line can be seen as a recording of the attack. So documentation purposes a report containing the data pre-processing process. The above information can be generated and used by the CF expert as potential legal digital evidence in the court of law.

4.7 User Interface

This proposed model separates the user interface from the data collection and processing elements. The administrator and computer forensic experts can communicate with the forensic logbook through user interface. Evidence Report can be generated after analysis.

During fusion at every level, in order to increase accuracy, external knowledge is always utilized as auxiliary information, and extracted information during processing procedure are saved to the knowledgebase making the process more dynamic.

4.8 Feasibility

The proposed model can be useful as an evidence acquisition tool for supplying the Offline Admissible Legal Digital Evidence for the Forensic Investigating agencies including preservation and continuity of evidence, and transparency of the CF methods. As the admissibility and weight are the two determinants in the legal acceptability of digital evidence [**6**], the courts deal with issues related to the difference between the novel scientific evidence and the legal evidence.

There are three requirements for the evidence to be admissible in the court [**2**]:

- Authentication (showing a true copy of the original)
- The best evidence rule (presenting the original)
- Exceptions to the hearsay rule. (allowable exceptions are when confession, business or other official records are involved)

From an evidence perspective, the law enforcement agencies will seek something that they can demonstrate to others long after the event is over (i.e. the evidence log file). The main aim is to identify the features that will be responsive to the needs of the law enforcement agencies in collecting the information and protecting the chain of evidence of computer intrusions, so that it will stand up in the court of law to prove the crime.

5. CONCLUSIONS AND FUTURE WORK

To collect the digital evidence is not an easy task. In this paper we have proposed a proprietary fusion based investigation model which can effectively process different types data both syntactically and semantically to retrieve the legal digital evidence. For tracking such types of coordinated multifaceted cyberspace attacks require cluster analysis techniques, adaptive neural networks, and rule-based knowledgebase systems.

Profiling, identifying, tracing, and apprehending cyber suspects are the important issues of research today. Within a computer system the anonymity afforded by the criminal encourages destructive behavior while making it extremely difficult to prove the identity of the criminal. Computer Forensics has emerged in response to the escalation of crimes committed by the use of computer systems either as an object of crime, an instrument used to commit a crime or a repository of evidence related to a crime. The evidence gathering process in a computing environment, by their nature is technical and different from other forms of evidence gathering. Data fusion along with data mining techniques promises to play a central role in the future prevention, detection, attribution, and remediation of such types of cyber crimes.

Our future work includes extending the investigation model to detect and prevent the various types of cyber threats. Furthermore, the co-author of this paper has also proposed a data fusion based dynamic risk analysis framework at the 2008 Society for Risk Analysis (SRA) annual conference proceeding symposium (Boston, MA, USA). In addition to the various military and non-military applications described in this paper, the application of these emerging data fusion methodologies can be effectively extended to related areas such as environmental and public health risk analysis, toxicology, and Health Care and Life Sciences (HCLS) data fusion. The methodology proposed here can find application in these emerging areas in terms of protecting environmental and human health ecosystems. Application of data fusion and data mash-up technologies facilitated by a semantic web informatics framework can increase the efficiency of dynamic data integration and risk analysis of various systems [**15**].

Disclaimer: Views expressed in this paper are those of authors' and do not necessarily represent: a) affiliating agency positions, or b) endorsement of specific tools.

6. REFERENCES

[1] E. Casey (ed.), **Handbook of Computer Crime Investigation**, Academic Press, 2001.

[2] E. Casey, **Digital Evidence and Computer Crime: Forensic Science**, Computer and the Internet, Academic Press, 2000.

[3] S Satpathy, A Kar, S Pradhan, **A Fusion based model for Computer Forensics**-Indian Science Congress conference-Jan3-7 2005.

[4] P.A. Porras. and P.G. Neumann, **EMERALD**: Event monitoring enabling responses to anomalous live disturbances in proceedings of the 20th National Information Systems Security Conference. National Institute of Standards and technology, 1997.

[5] H Lipson, **Tracking and Tracing Cyber Attacks: Technical Challenges and Global Policy Issues** (CMU/SEI-2002-SR-009), CERT Coordination Center, November 2002.

[6] J. Danielsson, **Project Description A system for collection and analysis of forensic evidence**, Application to NFR, April 2002.

[7] David L. Hall, Sonya A.H. McMullen, **Mathematical Techniques in Multisensor Data Fusion**,2nd edition, Artech House, 2004.

[8] D. Brezinski and T. Killalea, **Guidelines for Evidence Collection and Archiving**, RFC3227, February 2002.

[9] C. King, E. Osmanoglu, and C. Dalton, **Security Model Design, Deployment and Operations,** chapter 4, McGraw-Hill Osborne Media, 2001.

[10] P. Stephenson: **Intrusion Management: A Top Level Model for Securing Information Assets in an Enterprise Environment**, Proceedings of EICAR 2000, Brussels, Belgium, March 2000.

[11] http://www.data-fusion.org , accessed May 30, 2009.

[12] D. L. Hall, and J. Linas. **An Introduction to Multisensor Data Fusion**. In Proceedings of the IEEE, vol. 85, n° 1, pp. 6-23, 1997.

[13] P. Varshney, **Distributed Detection and Data Fusion**. Springer-Verlag, New York, NY., 1995

[14] E. Waltz and J. Linas, **Multisensor Data Fusion**. Artech House, Boston, MA, 1990.

[15] A. Mohapatra, **Semantic Web Informatics Facilitated Tool (SWIFT) – Dynamic Analysis of Risk Tools (DART): A Knowledgebase Framework**, Symposium on, "A Palette of Scientific Data - Online Tools to Support Risk Assessment", Society for Risk Analysis (SRA), Boston, MA, 2008.

[16] T. Bass, **Multi-sensor Data Fusion for Next Generation Distributed Intrusion Detection System**, In Proceedings of the IRIS National Symposium on Sensor and Data Fusion, 1999.