Automated Network Forensics

Laurence D. Merkle
Rose-Hulman Institute of Technology
5500 Wabash Ave., CM-103
Terre Haute, IN 47803
(812) 877-8474
l.merkle@ieee.org

ABSTRACT

The purpose of this research is to investigate the automated analysis of network based evidence in response to cyberspace attacks. The automated analysis techniques to be developed and studied will combine the efficiency of both existing and novel local search techniques with the scalability and robustness of evolutionary computation and other computational intelligence techniques.

Categories and Subject Descriptors

K.6.5 [**Management of Computing and Information Systems**]: Security and Protection – *unauthorized access.* I.2.8 [**Artificial Intelligence**]: Problem Solving, Control Methods, and Search – *heuristic methods.*

General Terms

Algorithms, Design, Experimentation, Human Factors, Legal Aspects, Security.

Keywords

Network forensics, computational intelligence.

1. INTRODUCTION

The purpose of this research is to investigate the automated analysis of network based evidence in response to cyberspace attacks. The automated analysis techniques to be developed and studied will combine the efficiency of both existing and novel local search techniques with the scalability and robustness of evolutionary computation and other computational intelligence techniques.

2. BACKGROUND

2.1 Network Forensics

Ad hoc work in computer forensics has been going on at least since the mid-1980's when a Lawrence Berkeley Laboratory astronomer named Cliff Stoll discovered a 75-cent error in a computer usage accounting program that eventually led him to a spy ring reporting to the KGB. However, serious theoretical work in the area did not begin until the Air Force Research Laboratory sponsored the First Digital Forensic Research Workshop. That meeting resulted in the following widely accepted definition of computer forensics:

The use of scientifically derived and proven methods toward the preservation collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations. [2]

Similarly, network forensics is defined to be

The use of scientifically proven techniques to collect, fuse, identify, examine, correlate, analyze, and document digital evidence from multiple, actively processing and transmitting digital sources for the purpose of uncovering facts related to the planned intent, or measured success of unauthorized activities meant to disrupt, corrupt, and compromise system components as well as providing information to assist in the response or recovery from these activities. [2]

As suggested by the definitions above, the goals of a forensic analysis vary, although service availability and mission continuity are always significant concerns. Stephenson suggests an outcome-oriented classification of goals: improved system defenses and accurately restored systems vs. legal or military action. [4] Other literature classifies goals according to the environment, each associated with emphasis on particular concerns: law enforcement, commercial, and military. [7] Obviously, these categorizations are not orthogonal.

In the law enforcement category, the primary concern is obtaining admissible evidence, necessitating the use of robust techniques that will withstand the scrutiny of the legal process. [6] In particular, the processes by which evidence is obtained and analyzed must be documented, reliable, repeatable, and explainable in terms understandable to the members of the court.

In the commercial and military categories, obtaining admissible evidence is less important than expeditiously identifying the means of attack in order to ensure service availability and mission continuity. In the military category, it is sometimes also desirable to determine the source of the attack. Given this information, a counterattack may be launched in order to maintain information superiority and continuity of mission-critical operations, as stated in the relatively new concept of Defensive Information Operations (DIO). [2] Effectiveness in this regard in certain situations demands the sacrifice of absolute certainty in the interest of speed.

In all cases, once an attack has been detected, the volume of data involved makes forensic analysis a challenging task. Carrier identifies two challenges of digital forensics: [1]

GECCO'08, July 12–16, 2008, Atlanta, Georgia, USA.

The Complexity Problem in digital forensics is that acquired data are typically at the lowest and most raw format, which is often too difficult for humans to understand.

The Quantity Problem in Digital Forensics is that the amount of data to analyze can be very large.

Both problems are exacerbated in the case of network forensics, in which the relevant data sources include various network traffic logs in addition to those of system events. Only after the data from these numerous sources are correlated can those that are clearly irrelevant to the attack be discarded. The remainder can then be analyzed to obtain the information necessary to achieve the organization's goals. Software tools exist that support this process by performing relevant statistical analyses (e.g. `tcpdstat`), scanning captured network traffic for intrusion detection system alerts (e.g. `snort`), analyzing network traffic sessions (e.g. `argus`, `tcptrace`), and viewing full content of individual sessions (e.g. `tcpflow`). As indispensable as these tools are, the parts of the problem that they solve are just the tip of the iceberg. Most of the analysis remains manual and time consuming.

Carrier identifies a number of open research questions that need to be addressed and tools that need to be developed based on that research to support all phases of this process. [2] For example, he cites the need for efficient filtering and intelligent data reduction techniques. In addition to their efficiency, he points out the need for the resulting technologies to be reliable, precise, accurate, non-reputable, secure, flexible, and inexpensive.

A few researchers have applied machine learning techniques in the context of network forensics. For example, Mukkamala and Sung have used support vector machines (SVMs) and artificial neural networks (ANNs) to which of 41 characteristics of TCP/IP network connections are the most reliable indicators of malicious activity. [5] Wei and Daniels have proposed the use of a hierarchical reasoning framework to generate evidence graphs for network forensics analysis. [8] Other researchers have recently suggested the need for automated problem solving techniques for improving the effectiveness and efficiency of network forensic analyses, mentioning various techniques including expert systems, search algorithms, machine learning, and other artificial intelligence techniques.

3. METHODOLOGY

A typical network forensics analysis proceeds in a series of stages, each of which provides clues that inform the investigator's efforts in the later stages [3]:

- Statistically unusual network traffic is identified using a tool such as `tcpdstat`. The resulting statistics often form the basis for the investigator's initial hypotheses regarding the system's compromise. For example, port scans and the use of unusual ports are immediately apparent at this stage.
- Alerts from tools such as `snort` are culled to identify those events that are potentially relevant to the compromise. The classifications attached to the alerts can serve to reinforce existing hypothesis. Other fields of the alerts often identify software used in the attack, network hosts and ports involved, and the approximate starting and ending times of various phases of the attack, for example. Alerts can also suggest additional hypotheses for investigation.
- Session data is extracted and analyzed statistically using a tool such as `Argus`. In the hands of a network forensics expert, the statistics resulting from this analysis serve as characteristic signatures of various network activities. For example, a series of relatively short sessions involving port 80 TCP is indicative of a scan for web server vulnerabilities, while a collection of two and three packet sessions involving a large number of ports clearly identifies port scanning activity.
- Full contents of individual packets are examined. The packet contents can confirm or contradict the hypotheses established in the earlier stages of analysis, and as such represent an essential link in the chain of evidence. The contents can also provide additional clues such as software "fingerprints" that can in turn lead to additional hypotheses.

Currently, the stages identified above are conducted independently. The use of information obtained in one stage to guide and inform actions in later stages requires human interpretation and explicit action. This research effort will integrate the tools for the various stages into a single system that exploits computational intelligence and other automation techniques to reduce the requirement for human intervention. In the remainder of this proposal, the resulting integrated tool is referred to as the automated network forensic tool.

The effort will progress in several steps:

- Build an isolated network of virtual machines, including a "honeynet" and an attacker system. Implement a hybrid evolutionary algorithm on the attacker system that generates reasonably realistic variations on known network attacks against the honeynet. Collect the resulting network forensic data for use in training the automated network forensic tool. Data collection will continue concurrently with the next several steps.
- Select a set of open source forensic analysis tools to use as the basis of the integrated system. The tools mentioned above are likely possibilities because of their popularity, but others will be considered based on their capabilities, performance, and existing application program interfaces (APIs).
- Formally characterize the information produced and required by each stage of the network forensics analysis, as well as the manual processes currently involved in transforming the information between its production in one stage and its use in another.
- Identify the most time-consuming and error-prone processes as candidates for automation. For example, because of the typically large number of false positives and high repetitiveness, manual analysis of alert data is both tedious and error prone.
- Design and prototype the integrated system. The design will include appropriate data structures for the exchange of information among the various analysis stages, and it will allow for each of the identified processes to operate in an appropriate subset of three modes: manual, partially automatic, and fully automatic. However, in the prototype all processes will operate only in the manual mode.
- Automate the previously identified time-consuming and error-prone processes, applying computational intelligence techniques as appropriate. Continuing the alert data example, both the time required and the likelihood of error

can be reduced by grouping the alerts based on their characteristics and their relationship to existing hypothesis. This is a straightforward instance of the clustering problem, which has been approached using a large variety of machine learning techniques, including evolutionary computation.

- Randomly partition the network forensic data sets generated by the virtual network into a training set, a component testing set, and an integration testing set. Following the computational learning theory techniques appropriate to the specific machine learning techniques chosen in the earlier steps, train each of the components of the automated network forensics tool against the training set.
- Evaluate each of the components against the component testing set. Evaluate the integrated system against the integration testing set, as well as against existing network forensics data sets.

4. DISCUSSION

This research will produce:

- A system for generating realistic network forensics datasets. This evolutionary algorithm-based system also will be useful in assessing the effectiveness against novel attacks of various security mechanisms, and in particular of intrusion detection systems.
- A formal characterization of the information required and the information generated by each of the stages of a standard network forensics investigation.
- An automated network forensics tool that integrates the standard tools used in each of the stages of a network forensics investigation. The tool will be open source and available to both network forensics researchers and practitioners. Furthermore, it will be extensible, thereby allowing other researchers to contribute to the further automation of the network forensics process.

5. REFERENCES

[1] Carrier, B. (2003). Defining Digital Forensic Examination and Analysis Tools Using Abstraction Layters. *International Journal of Digital Evidence , 1* (4).

[2] Digital Forensic Research Workshop. (2001). *Research Road Map*. Utica, NY.

[3] Jones, K. J., Bejtlich, R., & Rose, C. W. (2006). *Real Digital Forensics: Computer Security and Incident Response.* Addison-Wesley.

[4] Mocas, S. (2003). Building Theoretical Underpinnings for Digital Forensics. *Digital Forensice Workshop*. Cleveland, OH.

[5] Mukkamala, S., & Sung, A. H. (2003). Identifying Significant Features for Netowrk Forensic Analysis Using Artificial Intelligent Techniques. *International Journal of Digital Evidence , 1* (4).

[6] Noblett, M., Pollitt, M., & Presley, L. (2000). Recovering and Examining Computer Forensic Evidence. *Forensic Science Communications , 2* (4).

[7] Pollitt, M. (1995). Computer Forensics: an Approach to Evidence in Cyberspace. *Proceedings of the National Information Systems Security Conference*, (pp. 487-491). Baltimore, MD.

[8] Wang, W., & Daniels, T. E. (2005). Building Evidence Graphs for Network Forensics Analysis. *Proceedings of the 21st Annual Computer Security Applications Conference (ACSAC 2005).* IEEE Computer Society.