Global Information Assurance Certification Paper

Keith M. Dickter
GSEC Practical Requirement Version 1.4b

THE EVOLUTION OF DATA MINING AND RELATED SECURITY CORRELATION TECHNOLOGY

ABSTRACT

Data mining involves the discovery of relationships among fields of data databases. The basic concepts have evolved from early use in marketing, to its use in fraud detection and data security.

Recently, the basic ideas and technologies have evolved even further, as "security event management" products have emerged to help IT security personnel effectively deal with the huge volumes of available security-related information. These products can automatically correlate, and compare suspicious information gathered from different points in a computer system, in order to draw conclusions, and act upon, potential attacks and security violations.

This paper will discuss various security-related applications of data mining technology; the emergence of, and methodology behind, correlation and security event management technology; a security event management application and current security event products that are available.

DATA MINING

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. It consists of finding correlations or patterns among often dozens of fields in large databases. The basic technology used in data mining has existed for many years. One of the earliest applications of the technology was in the marketing field, where computers would obtain and analyze data from large databases. One of the most familiar such examples can be seen at supermarkets, where customers use "scan cards" during checkout. Information contained in these cards is then extracted and used to observe and analyze buying patterns in many different categories, for decision making, in order to increase profits.

The technology was taken a step further in the realm of fraud detection. Available software products use algorithms to determine patterns and relationships among data, which at the surface may often seem to be unrelated. They can then reveal cases that are exceptions to these norms or "rules", and which may be indicative of error or fraud. When extended to the field of information technology (IT) security, discrepancies or unusual activity may be the signs of hacker or system infiltrators.

The technology has been used in security-related applications by both the government and private sectors.

Data Mining in Government

The Federal Government has used data mining for such diverse purposes as customer relationship management, fraud detection and detecting and preventing terrorist activities.

Representative Curt Weldon (R-Pa.) was among those who wrote data mining integration proposals, years before the September 11th terror attacks [1]. Recently, he claimed that had his proposals been implemented, the attacks could have been prevented. He attributed this to the fact that there were 33 classified agencies in the Federal government and no means to connect all the available raw data. Weldon felt that an agency with centralized data-mining capability would have provided the information necessary to anticipate the attack.

In fact, the Defense Advanced Research Projects Agency (DARPA) in reaction to both the September 11 terror attacks, and the frequency of information technology changes, opened a new office focused on promoting the information technology aspects of national security [2]. The Informational Awareness Office was established with a mission to develop and exhibit technologies, such as data-mining products, capable of handling terrorist and other attacks.

DARPA is currently working on a controversial Total Information Awareness Program, which would create a huge database of information on Americans, including financial transactions and medical records. The database could then be mined to detect terrorist activity. The government expects to spend up to $575 million on such data mining projects between the years 2004 and 2007 [3].

Clementine is another data-mining tools used by government [4]. It is a data-mining tool that allows the users to compare data elements for individual records, with other records and then norm, and make predictions based on analysis and understanding of the relationships. Among other uses is the ability to detect fraud and reduce risk.

Applications of Data Mining in Security and Intrusion Detection

Data mining has also been used in the private sector to assist in setting and testing security devices and identifying possible intrusion attempts.

When a firewall for example, is originally set to specify whom to allow access to the network, the original assumptions may be erroneous. For example, some individuals who should be allowed access may be denied such access. By mining the access log, such errors may be revealed in the denial exceptions [5].

Intuitively, it would seem that data mining would be a very valuable tool in the area of intrusion detection. Since by definition, data mining does not require advance knowledge, such as information regarding existing threat methodologies, to perform its analyses, it may reveal previously unknown network attack types by showing conditions, which are contrary to the norm within the network. For example, it might indicate service disruptions occurring within specific intervals or large amounts of usage during periods that are normally quiet. These deviations can then be investigated by security personnel.

In fact, an experiment was conducted at Columbia University, to verify the value and usefulness of data mining [6]. A script was used to scan tcpdump data files and extract connection level information about the network traffic. For each Internet Control Protocol (TCP) packet the script processed packets between the ports of the participating hosts and checked such items such as proper 3-way handshake procedure and error recording, calculated connection statistics and observed the manner in which the connection was terminated (i.e. normal, abort etc.).

They took 80% of the data from the normal tcpdump and the remaining 20% contained both normal and embedded attack data. This was done five different times using a different 80% of the normal data each time. The connection data was run through the mining software. As expected, the intrusion data contained overall higher rates of deviations from the normal connection patterns than did the regular data, showing the value of data mining in detecting attempted attacks.

Another potential use for data mining products is analysis of access control utilities, such as IBM's RACF (Resource Access Control Facility). RACF contains information such as files, user profiles and access records. However, the security violations generated are lengthy and time-consuming to verify. As noted in an Institute of Internal Audit (IIA) article [7], audit software tools, such as ACL (Audit Command Language) can be very useful for analyzing these utilities. Information that can be derived includes revoked password information, non-expiring passwords, last logon dates (to indicate inactive users), IDs with various authorizations, emergency access IDs and installation data. Tests can also be run to determine the types of data to which an individual has access.

Primary advantages of ACL are its abilities with regard to defining files and its user-friendly flexibility for establishing the commands for each objective. However, the latter can sometimes be a "double edged sword", since initial knowledge of the inherent data relationships are necessary in order to choose the commands. This is where the ability of data mining to automatically, and

without bias, identify relationships and patterns, of which the auditor may unaware and would not have programmed into his ACL steps, can add considerable value.

These tools can work together in a complimentary fashion. ACL may be used to help define downloaded files, at which point the can be exported in the appropriate format to the data mining tool. Significant deviations or items differing from normal patterns that are revealed can then be the focus of ACL steps further quantification, summarization and investigation. The IIA author refers to a similar process when he notes that other software can be used to "cast a large net", after which ACL can focus on specific violations.

SECURITY EVENT CORRELATION AND MANAGEMENT

The problem

The quantities, types and levels of sophistication of threats to information security have been increasing substantially over the last several years. Moreover, businesses lose $6.6 million on average each time proprietary information is stolen [8]. At the same time, organizations had until recently been ill prepared to handles these malicious acts. Subsequently, however, organizations have become more security conscious and have implemented numerous security tools including, firewalls Virtual Private Networks (VPN's) and intrusion detection systems that can provide pertinent information that may be critical to mitigating or even averting attacks.

The attackers will often leave behind unique evidence in parts of the computer systems that they intend to victimize. This evidence, known as "events" can be found in the logs of the operating systems, servers, applications (for debugging functions), and security devices.

While all this information is extremely valuable, the sheer abundance of such data, as well as the variety of sources, has resulted in new difficulties. Today, when security teams try to discover attacks and unauthorized access by sifting through and trying to interpret tremendous amounts of raw data extracted from these logs, the process used is often highly inefficient. In fact, according to an article in CNN.com, between 60% and 90% of the time IT managers spend resolving problems is lost to diagnostics [9]. This comes at a time when they must deal with diminishing budgets and should ideally be maximizing the efficiency and effectiveness of existing resources. Worse, while security personnel may be preoccupied with voluminous false alarms, they may miss true security dangers.

Security management has, therefore, become much more complex because of the many parts of the network often involved simultaneously, and the need to connect them to follow the trail of the attack. Simple data mining technology alone would not always be sufficient to deal with such diverse data loads. The benefits of success though, are huge, as security personnel may be able to discover the attack and hopefully prevent costly harm from occurring. Thus, clearly, a more effective and efficient process was needed.

Correlation

Correlation, as it pertains to security, is the process of comparing data from multiple sources to determine patterns and relationships indicative of attacks and misuse. If security events could be easily correlated from the various locations noted above, and then consolidated, summarized and analyzed, this would be very valuable in simplifying and reducing the number of alarms and problems requiring investigation. According to the CNN.com article, correlation could, therefore, potentially reduce IT operational costs, as well as revenue lost to downtime, by many millions of dollars for large businesses [9].

However, the following impediments exist, as delineated by security, in its white paper [10]:

- Event data formats differ, complicating comparisons
- Data is stored in multiple locations (i.e. consoles and logs)
- Manual comparisons are very labor intensive
- When done manually, real time information is not provided
- Threats constantly change requiring the techniques to adapt

The Solution: Security Event Management

As noted above, security events are often complicated, apparently unrelated, originate from various sources in the computer system and need to be identified, related, prioritized and acted upon as soon as possible. Therefore, a new line of products called Security Event Management (SEM) tools has emerged. These products can correlate the many seemingly diverse events, automatically, in real time. The end result is that security personnel can now utilize their time more efficiently, by performing sophisticated investigative work using up-to-date already summarized data, rather than the more routine diagnostic tasks that they had to accomplish previously.

There are basically three stages required in the security event management cycle. The data must be prepared; the many isolated but related security events must be assembled to create one single relevant security incident, or “security

event chain", and the potential security impact to the organization and response must be determined.

The following steps regarding the mechanics of security event management were described in Matthew Caldwell's article appearing in the Information Systems Control Journal [11]:

Preparation

Data transport – data must be extracted in a timely manner, from the relevant security tools and brought to the automated system. Encryption and authentication are encouraged, to preserve the security and integrity of the data, are recommended.

Data normalization – The next step is to transform the data into a uniform format, ensuring that the data remains intact.

Data reduction – Unnecessary data should be eliminated to decrease the chance of errors, by compressing and filtering data and removing duplicates.

Creating the Event Chain

To begin it is best to use a basic correlation method to create the event chain including:

- Field correlation – the most basic type of correlation. Field correlation is comparing specific events to single or multiple fields in the normalized data. For example, a basic search across devices for TCP/IP service on port 80 (HTTP)

- Auto Correlation – An automatic method in which all fields are compared systematically for positive and negative correlations across one or multiple fields.

- OpenService, Inc. provides some examples of "Rule" type correlations of varying complexity [12]:

  - Comparing event data from various security events (such as FTP, Telnet, http) generated by one firewall residing on one server. The correlation software would check for conditional relationships among different kinds of logon activity. For example, six FTP login failures and four Telnet failures in a minute might be considered an indicator of hacking activity.

- Extracting event data from various security events with sources on a single security product, but which reside on multiple systems or servers.
  For example, correlation conditions would be checked for Telnet logon failures from one server firewall, FTP logon failures from a second server firewall, and hhtp logon failures from a third server firewall. When all three conditions exist, it might mean the organization is experiencing a distributed password attack, rather than mere cases isolated cases of forgotten passwords.

- Multiple events evaluated from multiple security products running on multiple servers, such as firewalls, intrusion detection and VPN (Virtual Private Network). Here the product may look for Telnet failures on one firewall server, TCP failures on a server with IDS and http failures on the VPN server.

Assessing Event Impact and Course of Action

Once the event is created, the potential impact of the event must be assessed and course of action determined [12]. For example:

- Child/Parent Interdependency – Ignores the less material results of certain events. For example, a rule could be established to say if 'A', 'B', 'C', and 'D' occur, report only 'A.' Thus, if the VPN and IDS are affected by a down firewall, the former issues can be ignored.

- Time-dependency – Includes rules to compare alarms and produce an action, such as "If 'A' occurs, followed by 'B' then perform 'C.

Relying solely on rules, however, could cause a decrease in performance as the amount of both data and rules expand and it is extremely difficult to establish all-inclusive rules when new attack methodologies arise almost daily.

Caldwell recommends using multiple data set techniques to determine the significance of events [11]. In this manner, false alarms will be minimized, and fewer true threats will be missed. Some examples are:

- Vulnerability correlation – The process of mapping IDS events that affect a host with the host's vulnerabilities. This method is effective because by assessing the chances of success for an attack, it is more likely that the attack will receive the correct level of priority and response.

- Open port correlation – Also, determines the probability of success of an attack, but by correlating it to the list of open port numbers on the targeted host being attacked. Events headed for an open destination port will be directed to a host that does not have this port open.

HIPAA – A Practical Application for SEM's

The 1996 Health Insurance Portability and Accountability Act (HIPAA) was a part of healthcare reform and patient's rights efforts. Contained within the legislation are efforts intended to simplify the processing of insurance claims and to promote electronic healthcare transactions for healthcare-related information using such factors as consistent diagnosis and transaction codes [13]. The grouping of entities subject to HIPAA is very extensive as they include not only health care providers of any size, but also, employers, health plans, government entities, colleges, vendors and any organization with access to patient data.

However, since increased computerization generally results in greater risks security exposures, the legislation included a comprehensive security subcategory, featuring five requirements designed to ensure that only patients, their healthcare professionals, and related necessary parties only, will have access to their medical records. :

- Administrative Procedures
- Physical Safeguards
- Technical Security Services
- Technical Security Mechanisms
- Electronic Signature

The "Technical Security Mechanisms Requirement" is especially relevant to the issues of SEM, as the regulations require organizations to develop procedures to prevent unauthorized access to data transmitted through a network. It specifies alarms, audit trails and event reporting as among the required safeguards.

Despite the specific references to technology objectives, HIPAA did not require specific compliance tools. Therefore, organizations have been utilizing numerous security products. However, HIPAA does require that organizations provide evidence that tools implemented are functioning as intended. In order to meet this requirement, organizations must extract, organize, analyze and maintain the significant security event data from these security products.

Thus, IT security personnel face a dilemma similar in nature to what was noted above with regard to intrusion detection. The security personnel are often

faced with huge amounts of varying data types, that they must analyze, act upon, and retain for reporting. The consequences of not doing so may be an indication that they are not adequately securing their data and complying with HIPAA. Currently, the overall target date for compliance by larger organizations is April 14, 2003; smaller organizations have until April 14, 2004. The legislation imposes heavy penalties for non-compliance. Moreover, there is the possibility of civil lawsuits against their organizations, should confidential data be stolen and even published.

A solution to this dilemma once again, is the use of an SEM tool, which has the ability to:

- Accumulate data from various security sources
- Determine patterns within these sources
- Establish priorities for security events
- Keep a security information database
- Allow for appropriate reporting

Current Products Available

There are a wide variety of SEM products available [14]. E-Security introduced the concept several years ago, and now finds itself with numerous competitors. However, the common element is the ability to monitor security devices from numerous vendors and normalize the data, aggregate the data and reduce the number of alarms and correlate alarms to prioritize significance.

Thus, while each product uses its own specific technologies to accomplish this, they all seek to identify security issues through patterns. The best products have built in rules, but have the flexibility to allow new rules to be written as relationships are revealed.

Following are some examples of available products:

Arcsight correlation system calculates “threat severity index”, which is based on an analysis of related events from various security tools, combined with vulnerability data for the targets and business information such as the value of the target items(s) to the organization. The index calculated would result in a range of pre-programmed possible actions available, which depend on the circumstances of the actual event.

For example, if a buffer overflow is detected by the IDS and the firewall determines that it has reached a vulnerable host, the product will find the source

and target address, scan the host for damage and block the source to prevent further damage [15].

NetForensics 3.0 takes security event information from many sources and combines them so that impending attacks can be investigated. The product allows for both wide views by IT staff at a network operations center, more localized views by administrators who need to prevent network threats. The tool normalizes the combined data, correlates events, and provides analysis and a real-time console by which to pinpoint specific security issues [16].

LogSmart by Network Intelligence, was created for use in high volume networks, and can work with data from 3,000 security tools, and analyze 60,000 events per second. Moreover, using compression, is can reduce the volume of event data by 95 percent [15].

SUMMARY

Data mining has been, and continues to be, a valuable technology resource. From its beginnings in the field of marketing to its expansion into the fields of fraud detection and security, it can reveal useful relationships and deviations worthy of follow-up and investigation. However, with vast quantities of network event data coming from diverse sources, the SEM products available today have the advantages of being able to normalize the data from multiple product types and brands, reduce the amounts of events and false alarms such that the significant issues can receive the appropriate focus; automate responses to attacks and provide all consolidated information in real time.

With all the new security equipment implemented as a result of attacks such as Code Red and Nimda, the decline in price and increase in available features presented by the SEM tools, it is expected that the market for these products will grow from $300 million this year to $600 million by the year 2006 [7]. The value of, and market for SEM tools will likely continue until such time as the next stage in the security evolutionary process is fully realized. According to Gartner, Inc., this is expected to be automated intrusion prevention products, which will eventually work within the network structure, utilizing the existing security devices to actually prevent the intrusions [17].

Endnotes

1. Verton, Dan

2. Caterinicchia, Dan (“Data Mining Aims . . .”)

3. Puzzanghera, Jim

4. Caterinicchia, Dan

5.Bhandari, Inderpal

6. Lee, Wenke and Salvatore J. Stolfo

7. Goldsmith, Jim

8. Hulme, George V

9. Drogseth, Dennis

10. e-security, inc

11. Caldwell, Matthew

12. Open Service, Inc

13. OpenService, Inc (Enabling HIPAA Compliance)

14. Desmond, Paul

15. Fisher, Dennis

16. Sturdevant, Cameron

17. Nicolett, Mark and Matthew Easley

References

Bhandari, Inderpal "Golden Means: Data mining and Security – The Last Line of Defense?" DSStar. July 28, 1998.
http://www.hpcwire.com/dsstar/98/0728/100241.html

Caldwell, Matthew. "The Importance of Event Correlation for Effective Security Management." Information Systems Control Journal, Volume 6, (2002): 36-38.

Caterinicchia, Dan. "SPSS Updates Data Mining Tool." Federal Computer Week. December 28, 2000.
http://www.fcw.com/fcw/articles/2000/1225/web-spss-12-28-00.asp

Caterinicchia, Dan. "Data mining Aims at National Security." Federal Computer Week. March 4, 2002.
http://www.fcw.com/fcw/articles/2002/0304/pol-darpa1-03-04-02.asp

Desmond, Paul. "Security Central." Network World Fusion. October 21, 2002.
http://www.nwfusion.com/supp/security2/savior.html

Drogseth, Dennis. "Taking Event Correlation Seriously." CNN.Com. March 20, 2000.
http://WWW.cnn.com/2000/TECH/computing/03/20/event.corr.idg/

e-security, inc. White Paper "Security Event Management: Correlation" October 7, 2002. Available by completing free form.
http://www.esecurityinc.com/productcorporateliterature/SEMCorrelationTechnology.pdf

Fisher, Dennis. "New Tools Automate Attack, Intrusion, Responses." eWeek. December 9, 2002.
http://www.eweek.com/print_article/0,3668,a=34536,00.asp

Goldsmith, Jim. "Perform an Audit on the RACF Database." Institute of Internal Auditors June 1, 2000.
http://www.theiia.org/itaudit/index.cfm?fuseaction=forum&fid=64

Hulme, George V. "Data Deluge." Information Week. August 19, 2002.
http://www.informationweek.com/story/IWK20020816S0036

Lee, Wenke and Salvatore J. Stolfo. "Data Mining Approaches for Intrusion Detection." Cs. Columbia.edu
http://www.cs.columbia.edu/~sal/hpapers/USENIX/usenix.html
Nicolett, mark and Matthew Easley. "The Emerging IT Security Management Market." Gartner, Inc. October 17, 2002.
http://www.dataquest.com/press_gartner/images/110845.pdf

Open Service, Inc. White Paper. "Enabling HIPAA Compliance With Open's Security Solutions." October 2002. Available by completing free form and selecting the download "Enabling HIPAA Compliance with SystemWatch." www.open.com

OpenService, Inc. White Paper. "Event Correlation: The Enabler of Active Internet Security Management." January,2002. http://www.itpapers.com/cgi/PSummaryIT.pl?paperid=24541&scid=286

Puzzanghera, Jim. "Senators Vow to Halt 'Data Mining' Project." Mercury News. January 17, 2002. http://www.siliconvalley.com/mld/siliconvalley/news/local/4969039.htm

Sturdevant, Cameron. "NetForensics Effectively Handles Hacks" EWeek. December 2, 2002. http://www.eweek.com/article2/0,3959,741464,00.asp

Verton, Dan. "Congressman Says Data Mining Could have Prevented 9-11." Computerworld. August 26, 2002. http://www.computerworld.com/databasetopics/data/story/0,10801,73773,00.html

Walker, John Q. "Security Event Correlation: Where Are We Now?" *NetIQ Corporation.* 2001. http://download.netiq.com/CMS/Security_Event_Correlation Where_Are_We_Now.pdf