ISSN:2249-5789
Anita Rajendra Zope et al , International Journal of Computer Science & Communication Networks,Vol 3(3),182-186

Event Correlation in Network Security to Reduce False Positive

Mrs.Anita Rajendra Zope, Prof. D.R.Ingle

***Abstract*—As the network based computer system plays an important role in modern society they have become target of our enemies and criminals. Therefore we need to find the best possible ways to protect our IT System. Different methods and algorithms are developed and proposed in recent years to improve intrusion detection systems. The most important issue in current systems is False Positive alarm rate. This is because current systems are poor at detecting novel anomaly attacks. These kinds of attacks refer to any action that significantly deviates from the normal behavior which is considered intrusion. Many NIDSs are signature based which consider only one device log, and conclude whether intrusion happened or not and internet attacks are increasing exponentially and there have been various attacks methods, consequently. This paper gives an overview of data mining field & security information event management system. We will see how various data mining techniques can be used in security information and event management system to enhance the capabilities of the system. we can use Data mining using Event Correlation Technique (ECT) for Network Intrusion Detection such thatby correlating events at different component of network security NIDS can identify whether actually intrusion occurred or not.**

***Index Terms*—Data mining, security information event management system.**

I. INTRODUCTION

In many industries computer network play an important role for information exchange, example tender quotations or for sending confidential information computer networks re most preferred. And so they have become the targets of our enemies and criminals. Therefore, we need to find the best ways possible to protect our systems. When Intrusion occurs security of system compromised. An intrusion can thus be defined as "any set of actions that attempt to compromise the integrity, confidentiality, or availability of a resource". Intrusion prevention techniques, such as user authentication (e.g., using passwords or biometrics), avoiding programming errors, and information protection (e.g., encryption) have been used to protect computer systems as a first line of defense. Firewall is one way to detect any malaises activity where the packets are analyses and discarded on the basis of the policies that are defined by the Network Administrator .But use of only firewall to defend is not good solution so the Intrusion Detection Systems IDS are now used for Intrusion detection rather than only firewall. Detection don't do anything to avoid or take any action to remove that problem but simple detect an intrusion and generate an alarm or sometime take first aid like block the IP address of the system from where any malicious packet came from. Data mining derives its name from the similarities between searching for gold in mines. Data mining is becoming increasingly common in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, and increase sales [2].If scope of data mining is applied to all events logs generated by various networking devices , system and application servers then efficiency of enterprise security can be drastically increased.

II. INTRUSION DETECTION SYSTEM

Intrusion Detection Systems (IDS) are security tools that, like other measures such as antivirus software, firewalls and access control schemes, are intended to strengthen the security of information and communication systems .It is a device or software application that monitors network and/or system activities for malicious activities or policy violations and produces reports to a Management Station [2]. Although intrusion detection technology is immature and should not be considered as a complete defense, but at the same time it can play a significant role in overall security architecture. If an organization chooses to deploy an IDS, a range of commercial and public domain products are available that offer varying deployment costs and potential to be effective. Because any deployment will incur ongoing operation and maintenance costs, the organization should consider the full IDS life cycle before making its choice. When an IDS is properly deployed, it can provide warnings indicating that a system is under attack, even if the system is not vulnerable to the specific attack [3]. These warnings can help users alter their installation's defensive posture to increase resistance to attack. In addition, IDS can serve to confirm secure configuration and operation of other security mechanisms such as firewalls. Within its limitations, it is useful as one portion of a defensive posture, but should not be relied upon as a sole means of protection. As e-commerce sites become attractive targets and the emphasis turns from break-ins to denials of service, the situation will likely worsen.

Host Based IDS(HIDS)

Its data come from the records of various host activities, including audit record of operation system, system logs, application pro-grams information, and so on.

Network Based IDS(NIDS)

A Network Intrusion Detection System (NIDS) is an intrusion detection system that tries to detect malicious activity such as denial of service attacks; port scans or even attempts to crack into computers by monitoring network traffic. A NIDS reads all the incoming packets and tries to find suspicious patterns known as signatures or rules. If, for example, a large number of

TCP connection re-quests to a very large number of different ports are observed, one could assume that there is someone conducting a port scan of some or all of the computer(s) in the network. Its data is mainly collected network generic stream going through network segments, such as: Internet packets [5] Another classification of intrusion detection is,

Anomaly Based IDS

Anomaly detection consists of first establishing the normal behavior profiles for users, programs, or other resources of interest in a system, and observing the actual activities as reported in the audit data to ultimately detect any significant deviations from these pro-files. Most anomaly detection approaches are statistical in nature. Anomaly detection systems can detect unknown intrusion since they require no a priori knowledge about specific intrusions. Statistical based approaches also have the added advantage of being adaptive to evolving user and system behavior since updating the statistical measures is relatively easy. Shortcomings of this type are , The selection of the right set of system (usage) features to be measured can vary greatly among different computing environments; The fine tuning of the deviation threshold is very ad hoc; User behavior can change dynamically and can be very inconsistent.

Existing NIDS

First major work in the area of intrusion detection was discussed by J.P Anderson. Concept that is introduced by that was as certain types of threats to the security of computer systems could be identified through a review of information contained in the system's log. This system log is available in many types of operating systems, particularly the various "flavors" of UNIX; automatically create a report which details the activity occurring on the system. Anderson identified three threats which could be identified from a concentrated review of the audit data:

- External Penetrations- Unauthorized users of the system
- Internal Penetrations- Authorized system users who utilize the system in an unauthorized manner.
- Misfeasors - Authorized user who mislead their access privileges [6]

Numbers of IDS are available in market, it's best to use a Web search to locate current products, reviews, and so forth. Commercial product literature is generally weighted towards marketing, which often makes it difficult to determine the product's functionality and detection approach. Virtually no commercial literature addresses issues such as the frequencies of false alarms, missed detections, or the system's sensitivity to traffic loads.

Shadow and Snort

Two public-domain ID tools, are unlikely to have the same level of support as commercial systems, so users will need a higher level of technical expertise to install and manage them. The effort involved is likely to pay off with a better understanding of ID and its strengths and limitations. Sensors usually reside at key monitoring points in the network, such as outside a firewall, while the analysis station resides inside the firewall. The sensor is based on public domain packet-capture software and does not pre-process the data, thus preventing an intruder from determining the detection objectives by capturing an unprotected sensor. Sensors extract packet headers and save them to a file that the analysis station reads periodically. The analysis station uses a Web-based inter-face to display filtering results as well as raw data. Shadow runs on many UNIX systems and Linux. Snort is a recent open-source public domain effort to build a lightweight, efficient, ID tool that can be deployed on a wide variety of UNIX platforms. According to the Snort Web site (www.snort.org), views are quickly outdated. The "Technology" sidebar describes a sample of commercial, research, and public domain tools. Snort is a lightweight network intrusion detection system, capable of performing real-time traffic analysis and packet logging on IP networks. It can perform protocol analysis, content searching/ matching and can be used to detect a variety of attacks and probes, such as buffer overflows, stealth port scans, CGI attacks, SMB probes, OS fingerprinting attempts, and much more. Snort uses a flexible rules language to describe traffic that it should collect or pass, as well as a detection engine that utilizes a modular plug-in architecture. Snort is currently undergoing rapid development. The user community is contributing auxiliary tools for analyzing and summarizing snort logs, providing additional capabilities. More importantly, there is a large group of users who contribute new signatures. As a result, new attacks are quickly represented in the signature database [3].

Problems in Current intrusion Detection System

We measure the quality of IDS by its effectiveness, adaptability and extensibility. An IDS is effective if it has both high intrusion detection (i.e., true positive) rate and low false alarm (i.e., false positive) rate. System is adaptable if it can detect slight variations of the known intrusions, and can be quickly updated to detect new intrusions soon after they are invented. It is extensible if it can incorporate new detection modules or can be customized according to (changed) network system configurations. Current IDSs lack effectiveness. The hand-crafted rules and patterns, and the statistical measures on selected system measures are the codified "expert knowledge" in security, system design, and the particular intrusion detection approaches in use. Expert knowledge is usually incomplete and imprecise due to the complexities of the network systems. Current IDSs also lack adaptability. Experts tend to focus on analyzing "current" (i.e., "known") intrusion methods and system vulnerabilities. As a result, IDSs may not be able to detect "future" (i.e., "unknown") attacks.

Large amount of Data

Another aspect which does not relate directly to misuse detection but is extremely important is how much data an analyst can efficiently analyze. That amount of data he needs to look at seems to be growing rapidly. Depending on the intrusion detection tools employed by a company and its size there is the possibility for logs to reach millions of records per day [8].

False positives: is the amount of false positives IDS will generate. A false positive occurs when normal attack is mistakenly classified as malicious and treated accordingly.

and knowledge. Consequently, data mining has enormous that human cannot process it fast enough to get the Data mining for security applications

In this section we will understand what is role of data mining in security information & event management system.

III. DATA MINING AND EVENT CORRELATION

SIEM system uses data feeds from various devices.

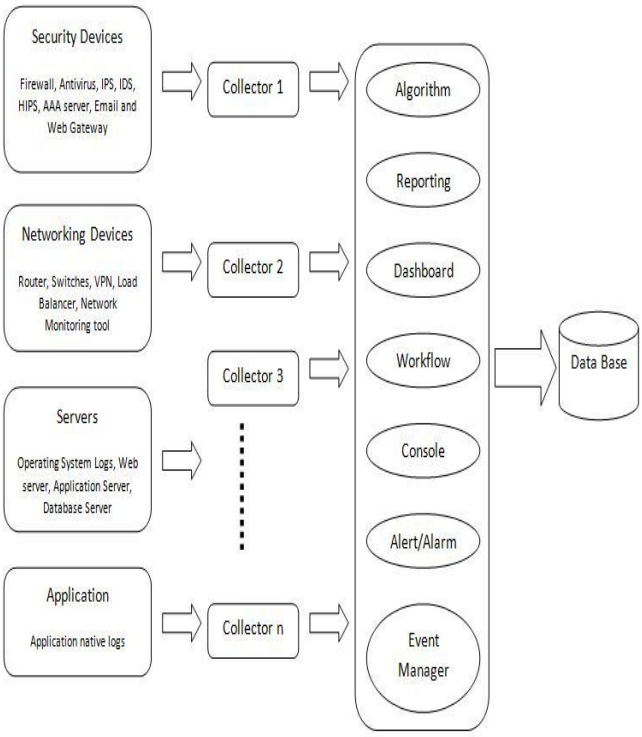

Fig. 1. SIEM Architecture

.

SIEM Architecture has four major parts:

1) *Data Sources :* SIEM system gets data feed from various devices which not only include networking devices but also some physical security devices like bio metric devices, card readers.
2) *Data Collectors:* primary function of data collector is to do normalization. This normalization happens in two ways it first normalize the values such as time zone, priority, severity in to common format, then they normalize the data structure in to common format. Some time collector do aggregation for example if there are 5 similar events in less than 3 second then collector can send only one such event. This filtering increases efficiency and accuracy and reduce processing time.
3) *Central Engine:* This is heart of SIEM system which mainly does applying data mining algorithm. This engine writes events in to database as they stream into the system. It simultaneously processes them through data mining engine where correlation happens. It also has user interface to display result of data mining algorithm. It enables end user to change certain properties of algorithm. Some of other component of this engine is reporting, alerting, and dashboards.
4) *Data Base:* As events stream in to central engine they are written in database with normalized schema. This storage helps us to do forensic analysis on historic data. By storing the events we can test new algorithm on historic data.

Data Mining Technology:

Data mining is, at its core, pattern finding. Data Mining is to extract knowledge interested by people from large database or data ware-house; the knowledge is implied, unknown and potentially useful information. Extracted knowledge is represented as concept, rule, law and model. The purpose of data mining is to help the decision-maker to find potential association between data, found neglected elements which are perhaps very useful for trends and decision-making behavior.

Common data mining methods and technologies are:

Correlation Analysis: correlation analysis was also called association rules, it is to find item set model knowledge frequently appeared from given data set, the purpose is to excavation the relationship that was hidden in data, for example, the customers that buy computer will buy some software, this is an association rules.

Sequential patterns: it is similar with correlation analysis, the purpose is also to excavate connection that between data, however, time series analysis focused more on the relationship of data in times, for example, and 80% people among printer buyer will buy printing paper after three months.

Classification: classification is to find model or function that can describe the typical characteristics of data set, so that it can identify ownership or categories of unknown data. Typical classify models have the linear regression model, the decision tree model, the model based on rule and the neural network model.

Clustering: Data was divided into a series of meaningful subset according to certain rules. In the same cluster, the gap between the individual is smaller, and in the different cluster, the gap is greater.

Deviation analysis: to find abnormal data from the database

Forecast: to find law according historical data, establish model, and to predict types, characteristics of the future data, etc based on the model.

Event correlation

Number of events happens in network in one second. An event in network management is typically defined as a piece of information dealing with a happening in the network, and may also be referred to as an alarm, due to its nature usually being something causing problems. Event correlation is defined in many different ways, but in its barest essence, an event correlator attempts to do exactly as the name suggests: associate events with one another in useful ways. Sometimes the sheer number of events which come in can be enough to overwhelm a engineer who cannot possibly treat each symptom separately. The object of event correlation is to attempt to pinpoint larger problems which could be causing many different symptoms to emerge. There are several subcategories of event correlation, including compression (duplication), count, suppression, and generalization. Compression reduces multiple occurrences of the same event into a single event, likely with some kind of counter. Count is

defined to be somewhat similar to compression: it is the substitution of a specified number of similar alarms with a single alarm. It is important to note that these need not necessarily be the same event, and also that there is a thresh-old associated with such a relation. Suppression associates a priority with alarms, and may choose to hide a lower priority alarm if a higher priority alarm exists. Finally, in the practice of generalization ion, alarms are associated with some sort of a superclass which is reported rather than the specific alarm. This could be seen to be useful to correlate events referring to multiple ports on the same switch or router if it has completely failed; it is unnecessary to see each particular failure if it can be determined that the entire unit is having problems.

Types of Event Correlation Systems

Rule-based Systems

A somewhat traditional approach to event correlation is that of rule-based analysis. In this approach, sets of rules are matched to events when they come in. Based on the results of each test, and the combination of events in the system, the rule-processing engine analyses data until it reaches a final state. It will then report a diagnosis, which could include, unfortunately, none at all, depending on the depth and capability of the rule set. Unfortunately, this approach does not necessarily perform well in respect to either of our criteria. For the results to be very accurate, an excessive amount of expert knowledge is typically needed to input the correct rules and keep them updated in case of any changes or new data. The rigidity of the path through the rule sets makes it so that events are may always be compared with an inordinate amount of test cases, slowing the system down and making correlation even more difficult.

Artificial Intelligence Systems

An approach that is radically different from the rule-based and codebook approaches uses various forms of artificial intelligence (AI). There are many different types of artificial intelligence, and event correlation techniques have been proposed which utilize various combinations of them, including Bayesian belief networks and expert systems. AI systems have an advantage in that, if well-programmed, they have the capability to be somewhat self-learning, helping to eliminate the continuous need for the expert knowledge of the previous systems. They also have the capability to sift through data at least as fast as the other systems to pro-duce their results. They claim that when they incorporate a tech-nique called inverse learning into their scheme, their system will never return a wrong answer as noise in the system increases[10].

Algorithms

Aggregated Risk and CALM algorithm

CALM (Compromise and Attack Level Monitor) is an assessment algorithm used by OSSIM to aggregate risk. Its input is a high volume of events, and its output is a single indicator of the general state of security of each asset.CALM algorithm displays in real time the compromise and attack level of networks.

• Compromise (C), measures the probability that an asset is compromised;

• Attack (A), measures how frequently an asset is being attacked.

Each asset has its A and C variables. Those are incremented according to three rules:

1. Any possible attack launched from machine 1 on machine 2 will increase the A of machine 2 and the C of machine 1.
2. When there is an attack response (an event that indicates the attack's success), the value of C will increase for both machines 1 and 2.
3. If the events are internal, the C value will go up only for the originating machine.

CALM is intended for real-time monitoring, hence, the algorithm should have a short-term memory that places importance on the most recent events and discards the oldest ones. Indeed, "Threshold" and "Recovery ratio" are two variables intended to periodically lower the C and A levels for each machine by a constant value.

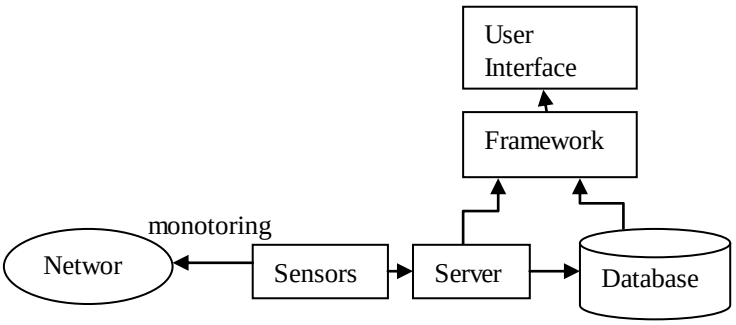

Fig 2. System architecture

Conclusion

This study shows that how data mining can be used in SIEM system. This paper firstly introduces the related knowledge, architecture of SIEM system and then the rule of algorithm for the correlation analysis. We have seen various association rules to detect abnormal patterns.

One of the areas we are exploring for future research is how we can use other data mining technique like classification, clustering to enhance the system capacity. In addition, we are enhancing the techniques we have mentioned to reduce false positive alerts and to reduce CPU load on system while computing data mining rules. Furthermore we are working to contribute some new modules for open source SIEM project.

REFERENCES

[1] I. K. Ravichandra Rao, "Data Mining and Clustering Techniques," *DRTC Workshop on Semantic Web*, DRTC, Bangalore , paper k , pp. 1-1, 8th – 10th December, 2003.

[2] Jeffrey W. Seifert, "Data Mining and Homeland Security:An Overview," *CRS Report* , pp. 1-1, Jan. 2007.

[3] Ming-Syan Chen,Jiawei Han,Philip , "Data Mining : An Overview from a Database Perspective," *IEEE Trans on knowledge and data engineering.*, vol. 8, no. 6, pp. 1-1, Dec 1996.

[4] Shuhong Yuan,Chijia Zou , "The Security Operations Center Based on Correlation Analysis "

[5] Entisar E. Eljadi, Zulaiha Ali Othman, "Anomaly Detection for PTM's Network Traffic Using Association Rule", *2011 3rd Conference on Data Mining and Optimization (DMO* ,June 2011

[6] Agrawal, R. & Srikant,,"Fast Algorithms for Mining Association Rules", Proceeding of the 20th VLDB Conference Santiago,1994

[7] Han, J., & Kamber, M. 2006. Data Mining Concepts and Techniques. Second Edition. The Morgan Kaufmann Series in Data Management Systems.
[8] Cynthia Rudin,Benjamin Letham,Ansaf Salleb-Aouissi , Eugene Kogan,David Madigan, "Sequential Event Prediction with Association Rules,"
[9] Theodoros Lappas, Konstantinos Pelechrinis Data Mining Techniques for (Network) Intrusion Detection Systems.
[10] Andreas M•uller, Christoph Goldi, Bernhard Plattner (2009) Event Correlation Engine.
[11] Michael Tiffany (2002) A Survey of Event Correlation Tech-niques and Related Topics.

Author's Profile

Mrs. Anita Rajendra Zope, Pursuing Master of Engineering in Computer Science from Mumbai University in M.G.M College of Engineering and Technology Navi Mumbai Currently working in St .John college of Engineering ,palghar.IEEE member.

Mr. D.R. Ingle (ISTE LM'2004) is Professor of Computer Engineering Department at Bharati Vidyapeeth College of Engineering, NaviMumbai, Maharashtra state, India received bachelor degree, and Master degree in computer engineering. He has Participated in more than 10 refresher courses to meet the needs of current technology. He has contributed more than 30 research papers at national, International Journals. He is life member of Indian Society of Technical Education. His area of interest are in Databases, intelligent Systems, and Web Engineering.