MINDS: A NEW APPROACH TO THE INFORMATION SECURITY PROCESS

E. E. Eilertson*, L. Ertoz, and V. Kumar
Army High Performance Computing Research Center
Minneapolis, MN 55414

K. S. Long
U.S. Army Research Laboratory
Adelphi, MD 20783

ABSTRACT

This paper describes the work the University of Minnesota is doing with the U.S. Army Research Laboratory to advance the state-of-the-art in network intrusion detection. The Minnesota Network Intrusion Detection System (MINDS) is a data mining based system for detecting unusual network behavior, and emerging cyber threats, *Ertoz et a.l 2004*. MINDS is enjoying great operational success in the ARL's Interrogator, *Long 2004*, information assurance architecture and at the University of Minnesota. MINDS routinely detects brand new attacks and other malicious behaviors which could not have been detected by signature based systems. In addition to detecting new attacks MINDS is very effective at discovering rogue communication channels and the exfiltration of data that are very difficult to identify with other tools.

1. Introduction

Advances in computer technology have brought new capabilities to the Army's warfighter, increased our battlefield supremacy, and sped up research and development. However, the technological advantages given to us by embracing computers can also be used by our adversaries as a force equalizer or multiplier. Adversaries realize that by breaking into the computer resources used to develop new battlefield technologies, they can steal the R&D the U.S. has been spending years developing, and put it to their own use, effectively neutralizing our research. In addition to compromising R&D systems, enemies know that it may be easier to compromise command and control via the network than by having a person physically infiltrate the organization. In a wartime situation this puts the lives and operations of all the allies in grave danger as the enemy may know about operations before the soldier on the ground does. Network intrusion detection is an important technology for ensuring cyber security.

The University of Minnesota is working with the U.S. Army Research Laboratory to advance the state-of-the-art in network intrusion detection. The Minnesota Network Intrusion Detection System (MINDS) is a data mining based system for detecting unusual network behavior, and emerging cyber threats. MINDS is enjoying great operational success in the ARL's Interrogator information assurance architecture and at the University of Minnesota. MINDS routinely detects brand new attacks which could not have been detected by signature based systems. In addition to detecting new attacks MINDS is very effective at discovering rogue communication channels and the exfiltration of data that are very difficult to identify with other tools.

The MINDS suite contains a variety of modules for collecting and analyzing massive amounts of network traffic. The input into the MINDS system can either be tcpdump data, netflow data, or data collected by our custom data collector. Typical analyses done inside of MINDS are scan detection, behavioral anomaly detection, clustering, summarization and communication pattern analysis. Independently, each of these modules provides key insights into the network and aids in identifying malicious behavior. When combined, as they automatically are inside of MINDS, these modules have a multiplicative affect on analysis.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE
00 DEC 2004

2. REPORT TYPE
N/A

3. DATES COVERED
-

4. TITLE AND SUBTITLE
Minds: A New Approach To The Information Security Process

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Army High Performance Computing Research Center Minneapolis, MN 55414; U.S. Army Research Laboratory Adelphi, MD 20783

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM001736, Proceedings for the Army Science Conference (24th) Held on 29 November - 2 December 2005 in Orlando, Florida. , The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**UU**	**4**	

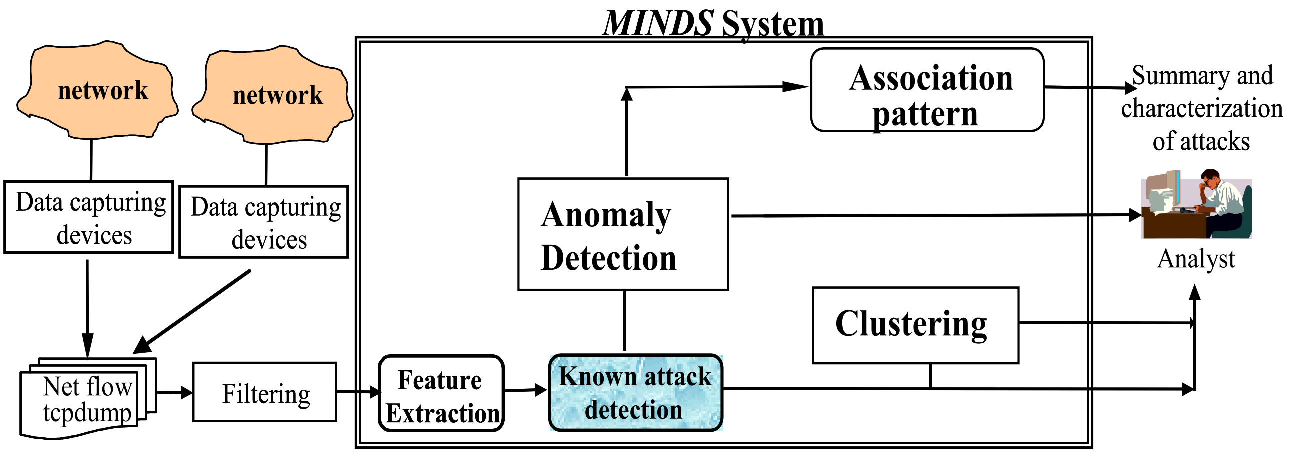

Figure 1 MINDS Architecture

The core of MINDS is a behavioral anomaly detection module based upon a revolutionary technique for calculating the distance/similarity between points in a high dimensional space. One of the key advantages of this technique is the ability to calculate similarity between categorical and numerical attributes. The new distance/similarity measure is incorporated in a new density based behavioral anomaly detection framework. Unlike other anomaly detection methods, this new framework does not suffer from a high number of false alarms. Combining this new way of measuring similarity between records with an advanced and very robust anomaly detection scheme has allowed us to develop the first effective anomaly detection scheme for intrusion detection. A key strength of this technique is its ability to find behavioral anomalies. Some real examples from its use in the DoD network are identification of streaming video from a DoD office to a computer in a foreign country and identification of a back door on a hacked computer. To the best of our knowledge, no other existing anomaly detection technique is capable of finding such complex behavior anomalies while maintaining very low false alarm rate.

In addition to the anomaly detection module, MINDS has a scan detection module to identify reconnaissance operations that are typically performed by attackers to identify vulnerable computers. Reliable detection of scans is important for securing the network perimeter, but existing schemes for scan detection are unable to detect slow and stealthy scans without unduly large false alarm rate. MINDS' scan detector is particularly suited for finding very slow and stealthy scans, and it has a nearly zero false alarm rate. A thorough evaluation at the University of Minnesota showed that of 2500 scans detected in 30 minutes at the University network border, containing 3.9 million connections, there were only 22 false alarms. At the ARL Center for Intrusion Monitoring and Protection (CIMP) the performance is even better, out of 1150 scans detected in one hour there were only 3 false alarms.

The MINDS suite not only aids in the detection of malicious behavior, but helps the security analyst better understand the network being monitored. One of the key modules for doing this is the clustering module. Using a combination of novel similarity/distance measures and a novel Shared Nearest Neighbor (SNN), *Ertoz et al. 2002, 2003*, clustering technique, the clustering module helps in identifying modes of behavior on the network, e.g. email traffic, web browsing of news sites, ftp traffic etc. Major modes usually reflect normal behavior, and minor modes often point towards malicious activity. Some rather benign modes detected in this way (from the data collected at ARL-CIMP) are peer-to-peer file sharing, and gaming. Some of the more dangerous modes detected in this way include backdoor communication channels and insiders scanning networks.

1.1 Scan (Surveillance) Detection

A precursor to many attacks on networks is often a reconnaissance operation, more commonly referred to as a scan. Identifying what attackers are scanning for can alert a system administrator or security analyst to what services or type of computers are being targeted and help them take preventative measures to protect the resources they oversee, e.g. installing patches, firewalling services from the outside, or removing services on machines which do not need to be running them. In addition, detection of outbound scans can help identify compromised machines within the network being protected. Despite the importance of scan detection, its value is somewhat overlooked in the security community, as most existing tools and techniques generate far too many false alarms, or are unable to detect stealthy scans targeted specifically at the monitored enterprise – the type of scans that analysts would really be interested in.

In our current research, we have developed a new scan detection technique that has much lower false alarm rate and much higher coverage than existing techniques, particularly suitable for detecting very slow stealthy scans. For example, an analysis of traffic data for a one-hour period from a DoD sensor resulted in the identification of 1,150 scans, with only 3 false alarms. In particular, the ability of our new schemes to detect very low volume scans has lead to the identification of several compromised computers used as stepping stones inside the Army's network, compromises that were not detected by any other method.

Our current research on scan detection is focusing on detecting scans that involve many scanners and multiple sites. Many sophisticated scans are performed in a distributed fashion – with the help of a large number of compromised computers that are used as scanners under the control of a single attacker but are located on physically and logically distributed sites. Each scanner is used to probe only a few computers at each site (to avoid detection) but collective scanning by many scanners can often provide the sophisticated attacker significant information about the various networks of interest. Detection of such distributed scans is particularly hard, as the activity of any single scanner at an individual site is hard to distinguish from normal traffic. To detect such sophisticated scans, we are developing new algorithms and tools for doing correlation of scans over time, and across sites. These tools will automatically identify scanners that are looking at networks at multiple sites, indicating a conscious effort to scan military sites as opposed to random scanning. A key challenge has been to develop algorithms that can efficiently identify groups of IP addresses that are collectively targeting non-trivial parts of the internal IP space on specific ports.

Scan detection can play a major role in second level analysis. Once a computer has been deemed compromised, a check of what IPs have scanned this computer, what services was it scanned for, and did it ever reply can give insight into how and when the computer was compromised. It can also guide the analyst to look for other computers compromised around the same time or manner. We are currently creating and testing tools for using scan detection in second level analysis. This has required putting all of the scan detection results into a database which may contain large amount of data, and creating a set of high-performance tools for quickly drilling down into the database to look for the relevant information.

2. PROFILING

Profiling techniques can play many useful roles in intrusion detection while augmenting the basic anomaly detector framework used in MINDS. Significant deviation from an established profile (of a service or a specific computer) can itself indicate abnormal (and thus potentially malicious) behavior. For example, one common technique hackers use to hide their activity is masquerading one service as another service, e.g. running an SSH service on port 80 which is used for web traffic. Such activity can be quickly detected if a detailed profile is available for each service which can be checked online for possible deviations. Identification of a change in the profile of a suspect computer could indicate the timeframe that the computer was compromised and allow an analyst to quickly understand the magnitude of damage possibly done, and look for other computers compromised in a similar way, or around the same time. In this project we have been developing profiling techniques for both services and individual IP addresses.

2.1 Profiling of Individual IP Addresses

A key requirement in the profiling of individual IP addresses is that the profiling programs need to be fast and interactive since these programs will be used for making time-critical decisions such as what needs to be done with compromised, potentially classified, computers. Pre-computation of profiles is not feasible due to the large number of IPs involved even in a moderate size network. In addition, the attributes of interest may be different for each specific investigation. We are developing HPC based solutions that can build individual profiles in an online and on-demand manner using long-term data available in archives. Since a typical analysis may require looking at a large number of IP addresses from multiple sites, there is an inherent concurrency in such computations that can allow effective use of HPC.

2.2 Profiling Services

A key challenge in development of profiles for different services is selection of appropriate features. Individual flow based features such as number of packets, number of bytes, histograms of byte patterns etc. are easy to compute, but have limited information. More useful features that involve flow sequences are much more complex and can be highly compute intensive, especially given the large volume of data. In this project we are creating solutions for generating service profiles that use both single-flow and multi-flow features.

3. LONG-TERM PATTERN ANALYSIS

Stealth attacks, especially those by sophisticated adversaries, can be difficult to identify as these attacks maintain a low profile and are specifically designed to evade detection by known IDS tools. Detection of these sophisticated attacks requires analysis that looks at data from multiple IDS tools and multiple sensors over a long period of time. In this project, we are developing and testing a framework for such analysis. This work is building upon our existing work on pattern analysis of long term network data, *Ertoz et al. 2003*.

Our recent work on a parallel formulation of the shared nearest neighbor clustering code for behavior modeling made it feasible to analyze a massive amount of network data using HPC computers. This approach allowed easy identification of minor and deviant modes of behavior, often relating to misconfigured computers, insider abuse, and policy violations that could not have been detected by other methods. These clusters can give analysts information they could act immediately on, as well as bring an understanding of the modes of behavior in the network traffic. While clustering is a very effective scheme for highlighting groups of similar connections, association patterns can be more effective for finding isolated patterns. In our future research, we will focus on two types of association patterns: support envelopes, *Steinbach et al. 2004*, and hypercliques, *Xiong et al. 2003*. They are particularly promising as they are not encumbered by a support threshold. Another potentially useful type of association analysis pattern is sequential pattern analysis. The goal of this type of analysis is to find rules that relate the occurrence of one set of events to the later occurrence of another set of events, e.g., a particular type of attack may be preceded by a particular pattern of network activity. Our future work in this area will investigate algorithms for computing sequential versions of hypercliques and support-envelop patterns in network traffic data.

Summarization of network data using association pattern analysis has been particularly useful in our current research, as a single pattern can represent a collection of connections that are part of the same larger attack. We are testing summarization schemes that can work with data from multiple sensors and multiple IDS tools. Key challenges being addressed are (i) how to deal with the heterogeneous nature of alarms created by different IDS tools; (ii) how to handle different levels of anomalies and alarms from different sensors.

CONCLUSIONS

In this paper we have discussed some of the challenges that need to be addressed in modern intrusion detection, and what steps we are taking to address these challenges. All of our research is very applied in nature, this is what has allowed us to have such great success in the real world, and able to meet the needs of the most demanding of clients, the U.S. Military and the modern warfighter.

REFERENCES

Ertoz, L., Steinbach, M, Kumar, V., 2002: A New Shared Nearest Neighbor Clustering Algorithm and its Applications, 2nd SIAM International Conference on Data Mining.

Ertoz, L., Steinbach, M, Kumar, V., 2003: Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data, 3rd SIAM International Conference on Data Mining.

Ertoz, L., Eilertson, E., Lazarevic, A., Tan, P. Kumar, V., Srivastava, J. Dokas, P., 2004: The MINDS – Minnesota Intrusion Detection System, AHPCRC Technical Report 2004-121

Ertoz, L., Steinbach, M, Kumar, V., 2003, Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data, AHPCRC Technical Report 2003-102

Ertoz, L., Eilertson, E., Lazarevic, A., Tan, P. Kumar, V., Srivastava, J. Dokas, P., 2003: Detection and Summarization of Novel Network Attacks using Data Mining, AHPCRC Technical Report 2004-108

Long, K. 2004: Interrogator Intrusion Detection Architecture, ARL-TR

Steinbach, M. Tan P., Xiong, H. and Kumar V., 2004: Generalizing the Notion of Support, AHPCRC Technical Report 2004-114

Steinbach, M. Tan P. and Kumar V., 2004: Support Envelopes: A Technique for Exploring the Structure of Association Patterns, AHPCRC Technical Report 2004-115

Xiong, H., Tan, P., Kumar, V., 2003: Mining Strong Affinity Association Patterns in Data Sets With Skewed Support Distribution, AHPCRC Technical Report 2003-122

This work was supported in part by the Army High Performance Computing Research Center (AHPCRC) under the auspices of the Department of the Army, Army Research Laboratory (ARL) under Cooperative Agreement number **DAAD19-01-2-0014.** *The content of which does not necessarily reflect the position or policy of the government, and no official endorsement should be inferred.*