2008 IEEE/IFIP International Conference on Embedded and Ubiquitous Computing

Data Mining for Security Applications

Bhavani Thuraisingham, Latifur Khan, Mohammad M. Masud, Kevin W. Hamlen
The University of Texas at Dallas
{bhavani.thuraisingham, lkhan, mehedy, hamlen}@utdallas.edu

Abstract

In this paper we discuss various data mining techniques that we have successfully applied for cyber security. These applications include but are not limited to malicious code detection by mining binary executables, network intrusion detection by mining network traffic, anomaly detection, and data stream mining. We summarize our achievements and current works at the University of Texas at Dallas on intrusion detection, and cyber-security research.

1. Introduction

Ensuring the integrity of computer networks, both in relation to security and with regard to the institutional life of the nation in general, is a growing concern. Security and defense networks, proprietary research, intellectual property, and data based market mechanisms that depend on unimpeded and undistorted access, can all be severely compromised by malicious intrusions. We need to find the best way to protect these systems. In addition we need techniques to detect security breaches.

Data mining has many applications in security including in national security (e.g., surveillance) as well as in cyber security (e.g., virus detection). The threats to national security include attacking buildings and destroying critical infrastructures such as power grids and telecommunication systems. Data mining techniques are being used to identify suspicious individuals and groups, and to discover which individuals and groups are capable of carrying out terrorist activities. Cyber security is concerned with protecting computer and network systems from corruption due to malicious software including Trojan horses and viruses. Data mining is also being applied to provide solutions such as intrusion detection and auditing. In this paper we will focus mainly on data mining for cyber security applications.

To understand the mechanisms to be applied to safeguard the nation's computers and networks, we need to understand the types of threats. In [1] we described real-time threats as well as non real-time threats. A real-time threat is a threat that must be acted upon within a limited time to prevent some catastrophic situation. Note that non real-time threats can become real-time threats as new information is uncovered. For example, one could suspect that a group of terrorists will eventually perform some act of terrorism. However, if subsequent intelligence reveals that this act will likely occur before July 1, 2008, then it becomes a real-time threat and we have to take actions immediately. If the time bounds are tighter such as "an attack will occur within two days" then we cannot afford to make any mistakes in our response.

There has been a lot of work on applying data mining for both national security and cyber security. Much of the focus of our previous paper was on applying data mining for national security [1]. In this part of the paper we will discuss data mining for cyber security. In section 2 we will discuss data mining for cyber security applications. In particular, we will discuss threats to computers and networks and describe applications of data mining to detect such threats and attacks. Some of our current research at the University of Texas at Dallas will be discussed in section 3. The paper is summarized in section 4.

2. Data Mining for Cyber Security

2.1. Overview

This section discusses information related terrorism. By information related terrorism we mean cyber-terrorism as well as security violations through access control and other means. Malicious software such as Trojan horses and viruses are also information related security violations, which we group into information related terrorism activities.

In the next few subsections we discuss various information related terrorist attacks. In section 2.2 we give an overview of cyber-terrorism and then discuss insider threats and external attacks. Malicious

978-0-7695-3492-3/08 $25.00 © 2008 IEEE
DOI 10.1109/EUC.2008.62

intrusions are the subject of section 2.3. Credit card and identity theft are discussed in section 2.4. Attacks on critical infrastructures are discussed in section 2.5. Data mining for cyber security is discussed in section 2.6.

2.2. Cyber-terrorism, Insider Threats, and External Attacks

Cyber-terrorism is one of the major terrorist threats posed to our nation today. As we have mentioned earlier, this threat is exacerbated by the vast quantities of information now available electronically and on the web. Attacks on our computers, networks, databases and the Internet infra-structure could be devastating to businesses. It is estimated that cyber-terrorism could cause billions of dollars to businesses. A classic example is that of a banking information system. If terrorists attack such a system and deplete accounts of funds, then the bank could loose millions and perhaps billions of dollars. By crippling the computer system millions of hours of productivity could be lost, which is ultimately equivalent to direct monetary loss. Even a simple power outage at work through some accident could cause several hours of productivity loss and as a result a major financial loss. Therefore it is critical that our information systems be secure. We discuss various types of cyber-terrorist attacks. One is the propagation of malicious mobile code that can damage or leak sensitive files or other data; another is intrusions upon computer networks.

Threats can occur from outside or from the inside of an organization. Outside attacks are attacks on computers from someone outside the organization. We hear of hackers breaking into computer systems and causing havoc within an organization. Some hackers spread viruses that damage files in various computer systems. But a more sinister problem is that of the insider threat. Insider threats are relatively well understood in the context of non-information related attacks, but information related insider threats are often overlooked or underestimated. People inside an organization who have studied the business' practices and procedures have an enormous advantage when developing schemes to cripple the organization's information assets. These people could be regular employees or even those working at computer centers. The problem is quite serious as some one may be masquerading as someone else and causing all kinds of damage. In the next few sections we will examine how data mining can be leveraged to detect and perhaps prevent such attacks.

2.3 Malicious Intrusions

Targets of malicious intrusions include networks, web clients and servers, databases, and operating systems. Many cyber-terrorism attacks are due to malicious intrusions. We hear much about of net-work intrusions. What happens here is that intruders try to tap into the networks and get the information that is being transmitted. These intruders may be human intruders or automated malicious software set up by humans. Intrusions can also target files instead of network communications. For example, an attacker can masquerade as a legitimate user and use their credentials to log in and access restricted files. Intrusions can also occur on databases. In this case the stolen credentials enable the attacker to pose queries such as SQL queries and access restricted data.

Essentially cyber-terrorism includes malicious intrusions as well as sabotage through malicious intrusions or otherwise. Cyber security consists of security mechanisms that attempt to provide solutions to cyber attacks or cyber terrorism. When discussing malicious intrusions or cyber attacks it is often helpful to draw analogies from the non cyber world—that is, non information related terrorism—and then translate those attacks to attacks on computers and networks. For example, a thief could enter a building through a trap door. In the same way, a computer intruder could enter the computer or network through some sort of a trap door that has been intentionally built by a malicious insider and left unattended perhaps through careless design. Another example is a thief's use of a stolen uniform to pass as a guard. The analogy here is an intruder masquerading as some-one else, legitimately entering the system and taking all the information assets. Money in the real world would translate to information assets in the cyber world. Thus, there are many parallels between non-information related attacks and information related attacks. We can proceed to develop counter-measures for both types of attacks.

2.4. Credit Card Fraud and Identity Theft

We are hearing a lot these days about credit card fraud and identity theft. In the case of credit card fraud, an attacker obtains a person's credit card and uses it to make unauthorized purchases. By the time the owner of the card becomes aware of the fraud, it may be too late to reverse the damage or apprehend the culprit. A similar problem occurs with telephone calling cards. In fact this type of attack has happened to me personally. Perhaps while I was making phone calls using my calling card at airports someone noticed the dial tones and reproduced them to make free calls. This was my

company calling card. Fortunately our telephone company detected the problem and informed my company. The problem was dealt with immediately.

A more serious theft is identity theft. Here one assumes the identity of another person by acquiring key personal information such as social security number, and uses that information to carry out transactions under the other person's name. Even a single such transaction, such as selling a house and depositing the income in a fraudulent bank account, can have devastating consequences for the victim. By the time the owner finds out it will be far too late. It is very likely that the owner may have lost millions of dollars due to the identity theft.

We need to explore the use of data mining both for credit card fraud detection as well as for identity theft. There have been some efforts on detecting credit card fraud (see [2]). We need to start working actively on detecting and preventing identity thefts.

2.5. Attacks on Critical Infrastructures

Attacks on critical infrastructures could cripple a nation and its economy. Infrastructure attacks include attacking the telecommunication lines, the electric, power, gas, reservoirs and water sup-plies, food supplies and other basic entities that are critical for the operation of a nation.

Attacks on critical infrastructures could occur during any type of attack whether they are non-information related, information related or bio-terrorism attacks. For example, one could attack the software that runs the telecommunications industry and close down all the telecommunication lines. Similarly, software that runs the power and gas supplies could be attacked. Attacks could also occur through bombs and explosives. That is, the telecommunication lines could be physically attacked. Attacking transportation lines such as highways and railway tracks are also attacks on infrastructures.

Infrastructures could also be attacked by natural disaster such as hurricanes and earth quakes. Our main interest here is the attacks on infrastructures through malicious attacks, both information related and non-information related. Our goal is to examine data mining and related data management technologies to detect and prevent such infrastructure attacks.

2.6. Data Mining for Cyber Security

Data mining is being applied to problems such as intrusion detection and auditing. For example, anomaly detection techniques could be used to detect unusual patterns and behaviors. Link analysis may be used to trace self-propagating malicious code to its authors. Classification may be used to group various cyber attacks and then use the profiles to detect an attack when it occurs. Prediction may be used to determine potential future attacks depending in a way on information learnt about terrorists through email and phone conversations. Also, for some threats non real-time data mining may suffice while for certain other threats such as for network intrusions we may need real-time data mining. Many researchers are investigating the use of data mining for intrusion detection. While we need some form of real-time data mining, that is, the results have to be generated in real-time, we also need to build models in real-time. For example, credit card fraud detection is a form of real-time processing. However, here models are usually built ahead of time. Building models in real-time remains a challenge. Data mining can also be used for analyzing web logs as well as analyzing the audit trails. Based on the results of the data mining tool, one can then determine whether any unauthorized intrusions have occurred and/or whether any unauthorized queries have been posed.

Other applications of data mining for cyber security include analyzing the audit data. One could build a repository or a warehouse containing the audit data and then conduct an analysis using various data mining tools to see if there are potential anomalies. For example, there could be a situation where a certain user group may access the database between 3 and 5am in the morning. It could be that this group is working the night shift in which case there may be a valid explanation. However if this group is working between say 9am and 5pm, then this may be an unusual occurrence. Another example is when a person accesses the databases always between 1 and 2pm; but for the last 2 days he has been accessing the database between 1 and 2am. This could then be flagged as an unusual pattern that would need further investigation.
Insider threat analysis is also a problem both from a national security as well from a cyber security perspective. That is, those working in a corporation who are considered to be trusted could commit espionage. Similarly those with proper access to the computer system could plant Trojan horses and viruses. Catching such terrorists is far more difficult than catching terrorists outside of an organization. One may need to monitor the access patterns of all the individuals of a corporation even if they are system administrators to see whether they are carrying out cyber-terrorism activities [3], [4].

While data mining can be used to detect and prevent cyber attacks, data mining also exacerbates

some security problems such as inference and privacy. With data mining techniques one could infer sensitive associations from the legitimate responses. For more details on privacy we refer to [5], [6].

3. Our Current Research and Development

3.1 Data Mining for Intrusion and Malicious Code Detection

We are developing a number of tools that use data mining for cyber security applications at the University of Texas at Dallas, including tools for intrusion detection, malicious code detection, and botnet detection. An intrusion can be defined as any set of actions that attempts to compromise the integrity, confidentiality, or availability of a resource. As systems become more complex, there are always exploitable weaknesses due to design and programming errors, or through the use of various "socially engineered" penetration techniques. Computer attacks are split into two categories, host-based attacks and network based attacks. Host-based attacks target a machine and try to gain access to privileged services or resources on that machine. Host-based detection usually uses routines to obtain system call data from an audit-process which tracks all system calls made by each user-process.

Network-based attacks make it difficult for legitimate users to access various network services by purposely occupying or sabotaging network resources and services. This can be done by sending large amounts of network traffic, exploiting well-known faults in networking services, overloading network hosts, etc. Network-based attack detection uses network traffic data (i.e., tcpdump) to look at traffic addressed to the machines being monitored. Intrusion detection systems are split into two groups: anomaly detection systems and misuse detection systems.

Anomaly detection is the attempt to identify malicious traffic based on deviations from established normal network traffic patterns. Misuse detection is the ability to identify intrusions based on a known pattern for the malicious activity. These known patterns are referred to as signatures. Anomaly detection is capable of catching new attacks. However, new legitimate behavior can also be falsely identified as an attack, resulting in a false positive. The focus with the current state of the art is to reduce false negative and false positive rate.

We have used multiple models such as support vector machines (SVM). However we have improved SVM a great deal by combining it with a novel algorithm that we have developed. We will describe this novel algorithm as well as our approach to combining it with SVM. In addition we will also discuss our experimental results. For more details of our research we refer to [7].

Our other tools include those for email worm detection, malicious code detection, buffer overflow detection, botnet detection, and analysis of firewall policy rules. For email worm detection we ex-amine emails and extract features such as "number of attachments" and the train a data mining tools with techniques such as SVM and Naïve Bayesian classifiers to develop a model. Then we test the model to determine whether the email has a virus/worm. We use training and testing data sets posted on various web sites [8]. For firewall policy rule analysis we use association rule mining techniques to determine whether there are any anomalies in the policy rule set [9].

Similarly, for malicious code detection we extract n-gram features both with assembly code and binary code. We train the data mining tool with SVM and then test the model. The classifier then predicts whether the code is malicious. For buffer overflow detection we assume that malicious messages contain code while normal messages contain data. Distinguishing code from data is difficult on many computing architectures such as Windows x86 architectures because of variable-length instruction encodings, mixtures of code and data in each segment of the binary, and encrypted or compressed code segments. While these obstacles have impeded standard disassembly-based static analyses, we have found success using SVM training and testing [10].

3.2. Data Mining for Botnet Detection

Our current research with the University of Illinois Urbana Champaign is focusing in applying data mining techniques for *botnet* detection. The term "bot" comes from the word robot. A bot is typically autonomous software capable of performing certain functions. A botnet is a network of bots that are used by a human operator or botmaster to carry out malicious actions. Botnets are one of the most powerful tools used in cyber-crime today, being capable of effecting distributed denial-of-service attacks, phishing, spamming, and eavesdropping on remote computers. Often businesses, governments, and individuals are facing million-dollar damages caused by hackers with the help of botnets. It is a major challenge to the cyber-security research community to combat this threat.

Botnets have different topologies and protocols. The most prevalent botnets use communications based on Internet Relay Chat (IRC), and have a centralized architecture. There are many approaches available to detect and dismantle these IRC botnets. On the other hand, Peer-to-Peer (P2P) networks are a relatively new technology used in botnets. P2P botnets use decentralized P2P protocols to communicate among the bots and the botmaster. These botnets are distributed, having no central point of failure. As a result, these botnets are more difficult to detect and destroy than the IRC botnets. Moreover, most of the current research related to P2P botnets are in the analysis phase. The main goal of our project is to devise an efficient technique to detect P2P botnets. We approach this problem from a data mining perspective. We are developing techniques to mine net-work traffic for detecting P2P botnet traffic.

Our research on the botnet problem follows from the important observation that network traffic (as well as botnet traffic) is a continuous flow of data stream. Conventional data mining techniques are not directly applicable to stream data because of concept drift and infinite-length. We propose a technique that can efficiently handle both problems. Our main focus is to adapt three major data mining techniques: classification, clustering, and outlier detection to handle stream data. Our preliminary study on the development of new stream classification techniques for P2P botnet detection has encouraging results. [11]

4. Summary and Directions

This paper has discussed data mining for security applications. We first started with a discussion of data mining for cyber security applications and then provided a brief overview of the tools we are developing. Data mining for national security as well as for cyber security is a very active research area. Various data mining techniques including link analysis and association rule mining are being explored to detect abnormal patterns. Because of data mining, users can now make all kinds of correlations. This also raises privacy concerns.

One of the areas we are exploring for future research is active defense. Here we are investigating ways to monitor the adversaries. For such monitoring to be effective, the monitor must avoid detection by the static and dynamic analyses employed by standard anti-malware packages. We are therefore developing techniques that can dynamically adapt to new detection strategies and continue to monitor the adversary. We are exploring the use of adaptive machine learning techniques for this purpose. In addition, we are enhancing the techniques we have developed to reduce false positive and false negatives. Furthermore, we are exploring the applicability of our techniques to distributed and pervasive environments.

5. References

[1] Thuraisingham, B., "Web Data Mining Technologies and Their Applications in Business Intelligence and Counter-terrorism", *CRC Press*, FL, 2003.

[2] Chan, P, et al, "Distributed Data Mining in Credit Card Fraud Detection", *IEEE Intelligent Systems*, 14 (6), 1999.

[3] Lazarevic, A., et al., "Data Mining for Computer Security Applications", Tutorial *Proc. IEEE Data Mining Conference*, 2003.

[4] Thuraisingham, B., "Managing Threats to Web Databases and Cyber Systems, Issues, Solutions and Challenges", *Kluwer*, MA 2004 (Editors: V. Kumar et al).

[5] Thuraisingham B., "Database and Applications Security", *CRC Press*, 2005.

[6] Thuraisingham B., "Data Miming, Privacy, Civil Liberties and National Security", *SIGKDD Explorations*, 2002.

[7] Khan, L., Awad, M. and Thuraisingham, B. "A New Intrusion Detection System using Support Vector Machines and Hierarchical Clustering", *The VLDB Journal:* ACM/Springer-Verlag, 16(1), page 507-521, 2007.

[8] Masud, M. M., Khan, L. and Thuraisingham, B. "Feature based Techniques for Auto-detection of Novel Email Worms", In *Proc. 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2007)*, Nanjing, China, May 2007, page 205-216.

[9] Abedin, M., Nessa, S., Khan, L., Thuraisingham, B., "Detection and Resolution of Anomalies in Firewall Policy Rules", In *Proc. 20th IFIP WG 11.3 Working Conference on Data and Applications Security (DBSec 2006),* Springer-Verlag, July 2006, Sophia Antipolis, France, page 15-29.

[10] Masud, M. M., Khan, L, Thuraisingham, B., Wang, X., Liu, P., and Zhu, S., "A Data Mining Technique to Detect Remote Exploits", In *Proc. IFIP WG 11.9 International Conference on Digital Forensics*, Japan, Jan 27-30, 2008.

[11] Masud, M. M., Gao, J., Khan, L., Han, J., Thuraisingham, B., "Peer to Peer Botnet Detection for Cyber-Security: A Data Mining Approach". In *Proc. Cyber Security and Information Intelligence Research Workshop (CSIIRW 08)*, Oak Ridge National Laboratory, Oak Ridge, TN, May 12-14, 2008.