A Study on Data Mining and Knowledge Discovery in Cyber Security

Ashish Kumar Nayyar*

Abstract

In twenty first century, computing and the use of Internet have become omnipresent across the globe. The Internet is everywhere we cannot walk more than 20 feet in an urban area without finding an open connection. 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. This paper is a study on Data Mining, Cyber Security and various Data Mining Techniques that are being used for cyber security.

Keywords: Databases, Data Mining, Discovery, Knowledge Database, KDD system, Machine Learning, Cyber Security, Intrusion

Introduction

In the era of information society, computer networks and their related applications are becoming more and more popular. To defend against various cyber attacks and computer viruses, lots of computer security techniques have been intensively studied in the last decade, namely cryptography, firewalls, anomaly and intrusion detection. Among them, network intrusion detection (NID) has been considered to be one of the most promising methods for defending complex and dynamic intrusion behaviors.

Intrusion Detection is a process of monitoring and analyzing the activities in the computer system. The main objective is to identify the threats to the system, and then to protect and safeguard the system from those threats. Some of the technologies have has been verified and applied [11].

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. There has been a lot of work on applying data mining for both national security and cyber security. [21]

Ashish Kumar Nayyar*
IITM, GGSIPU

Data Mining

Data mining is a logical process that is used to search through large amounts of information in order to find important data. The goal of this technique is to find patterns that were previously unknown. Once you have found these patterns, you can use them to solve a number of problems. Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cut costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified.

Data Mining and Knowledge Discovery in Database

Historically, the notion of finding useful patterns in data has been given a variety of names, including data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing. The term *data mining* has mostly been used by statisticians, data analysts, and the management information systems (MIS) communities. It has also gained popularity in the database field. The phrase **knowledge discovery in databases** was coined at the first KDD workshop in 1989 (Piatetsky-Shapiro 1991)[2] to emphasize that knowledge is the end product of a data-driven discovery. It has been popularized in the AI and machine-learning fields. In my view, KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. *Data mining* is the application of specific algorithms for extracting patterns from data. The distinction between the KDD process and the data-mining step (within the process) is a central point of this article. The additional steps in the KDD process, such as data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge, and proper interpretation of the results of mining are essential to ensure that useful knowledge is derived from the data. Blind application of data-mining methods (rightly criticized as data dredging in the statistical literature) can be a dangerous activity, easily leading to the discovery of meaningless and invalid patterns.

The Six-step Knowledge Discovery And Data Mining Process

The goal of designing a Data Mining & Knowledge Discovery process model is to come up with a set of processing steps to be followed by practitioners when they execute their Data Mining & Knowledge Discovery projects. Such process model should help to plan, work through, and reduce the cost of any given project by detailing procedures to be performed in each of the steps. The Data Mining & Knowledge Discovery process model should provide a complete description of all the steps from problem specification to deployment of the results.

The six-step DMKD process [4] is described as follows:

1. **Understanding the problem domain**
   In this step one works closely with domain experts to define the problem and determine the project goals, identify key people and learn about current solutions to the problem. It involves learning domain-specific terminology.

2. **Understanding the data**
   This step includes collection of sample data, and deciding which data will be needed including its format and size. If background knowledge does exist some attributes may be ranked as more important. Next, we need to verify usefulness of the data in respect to the DMKD goals.

3. **Preparation of the data**
   This is the key step upon which the success of the entire knowledge discovery process depends; it usually consumes about half of the entire project effort. In this step, we decide which data will be used as input for data mining tools of step 4. It may involve sampling of data, running correlation and significance tests, data cleaning like checking completeness of data records, removing or correcting for noise, etc. The cleaned data can be further processed by feature selection and extraction algorithms (to reduce dimensionality), by derivation of new attributes

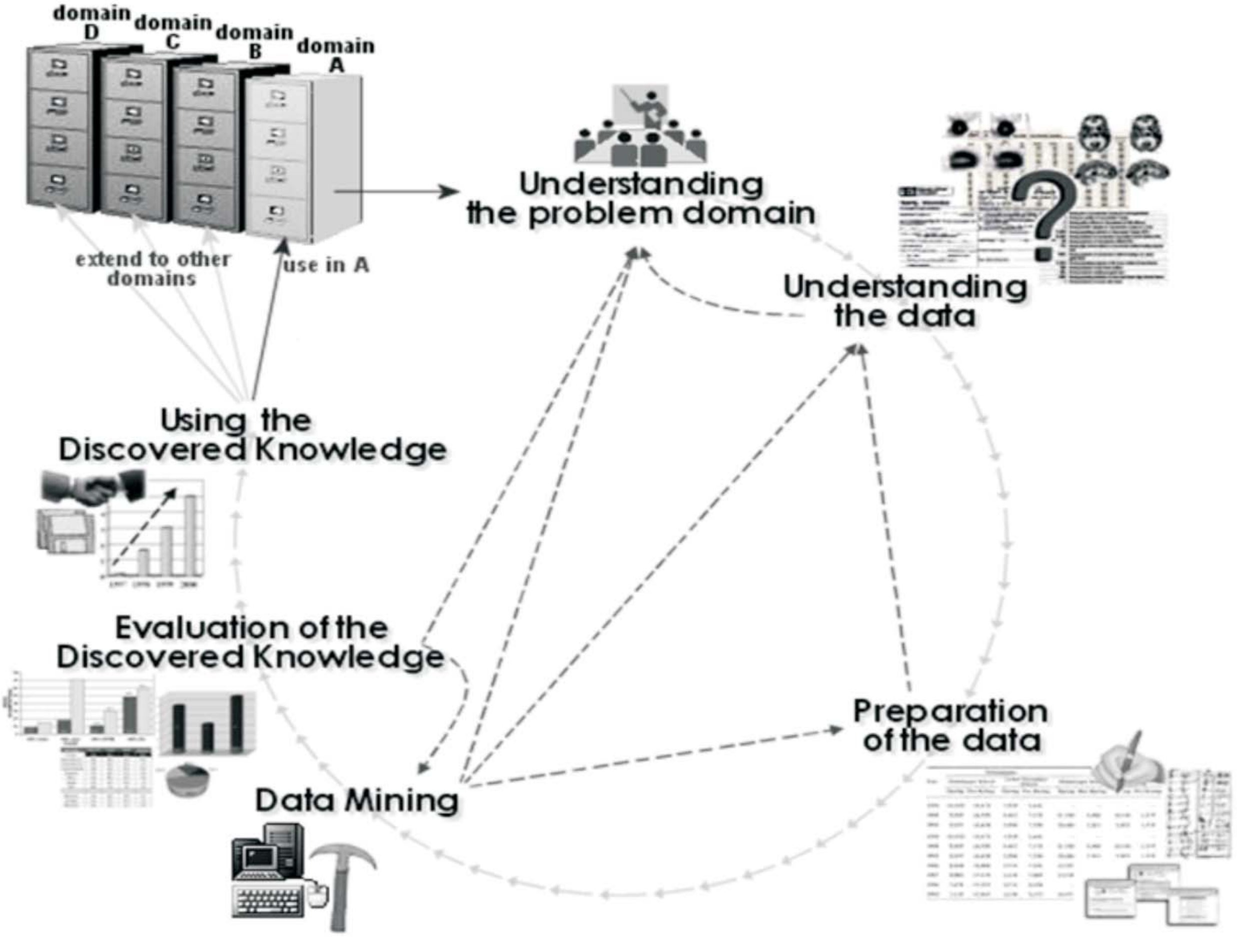

Fig. 1: The six-step DMKD process model [4]

(say by discretization), and by summarization of data (data granularization). The result would be new data records, meeting specific input requirements for the planned to be used DM tools.

4. **Data mining**
   This is another key step in the knowledge discovery process. Although it is the data mining tools that discover new information, their application usually takes less time than data preparation. This step involves usage of the planned data mining tools and selection of the new ones. Data mining tools include many types of algorithms, such as rough and fuzzy sets, Bayesian methods, evolutionary computing, machine learning, neural networks, clustering, preprocessing techniques, etc.

5. **Evaluation of the discovered knowledge**
   This step includes understanding the results, checking whether the new information is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. Only the approved models (results of applying many data mining tools) are retained. The entire Data Mining & Knowledge Discovery process may be revisited to identify which alternative actions could have been taken to improve the results. A list of errors made in the process is prepared.

6. **Using the discovered knowledge**
   This step is entirely in the hands of the owner of the database. It consists of planning where and how the discovered knowledge will be used. The application area in the current domain

should be extended to other domains. A plan to monitor the implementation of the discovered Knowledge should be created, and the entire project documented.

The DMKD process model just described is visualized in Figure-1. The important issues are the iterative and interactive aspects of the process. Since any changes and decisions made in one of the steps can result in changes in later steps, the feedback loops are necessary. The model identifies several such feedback mechanisms:

- From Step 2 to Step 1 because additional domain knowledge may be needed to better understand the data
- From Step 3 to Step 2 because additional or more specific information about the data may be needed before choosing specific data preprocessing algorithms (for instance data transformation or discretization)
- From Step 4 to Step 1 when the selected DM tools do not generate satisfactory results, and thus the project goals must be modified
- From Step 4 to Step 2 in a situation when data was misinterpreted causing the failure of a DM tool (e.g. data was misrecognized as continuous and discretized in Step 3). The most common scenario is when it is unclear which DM tool should be used because of poor understanding of the data.
- From Step 4 to Step 3 to improve data preparation because of the specific requirements of the used DM tool, which may have not been not known during the data preparation step.
- From Step 5 to Step 1 when the discovered knowledge is not valid. There are several possible sources of such a situation: incorrect understanding or interpretation of the domain, incorrect design or understanding of problem restrictions, requirements, or goals. In these cases the entire DMKD process needs to be repeated.
- From Step 5 to Step 4 when the discovered knowledge is not novel/interesting/useful. In this case, we may choose different DM tools and repeat Step 4 to extract new and potentially novel, interesting, and thus useful knowledge.

Cyber Security Terrorism

Overview

Cyber Security Terrorism relates to security violations through access control and various other means. Malicious software such as Trojan horses and viruses are other kinds of Cyber Security Terrorism. The next section, discusses various cyber related terrorist attacks like Intrusion, Credit Cards and Identity thefts, etc.

Cyber-terrorism-Internal and External

Among various national level terrorism risks one of the major threats is of Cyber-terrorism. Reason for this problem is availability of 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 lose of rupees. By disrupting 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.

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 someone may be masquerading as someone else and causing all kinds of damage.

Network Intrusions

Problem of 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 network 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.

Credit Card Fraud and Identity Theft

The most common form of attack is theft of credit card details. Here an attacker tries to steal our credit card information and 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. 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.

Data Mining Techniques For Intrusion Detection

Data Mining 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. A preliminary study on the development of new stream classification techniques for P2P botnet detection has shown encouraging results[12].

Zeus is a Trojan horse for Windows that was created to steal bank information using botnets. First discovered in 2007, Zeus spread through email, downloads, and online messaging to users across the globe. Zeus botnets used millions of zombie computers to execute keystroke logging and form grabbing attacks that targeted bank data, account logins, and private user data. The information gathered by Zeus botnets has been used in thousands of cases of online identity theft, credit card theft, and more[17].

Data mining based passive analysis to identify botnet traffic. Their approach is based on correlating multiple log files obtained from different points of the network. The system is not only to detect IRC-based botnet but also applicable for non-IRC botnets. The method is also effective because of its passive and regardless of payload nature. Hence, it is applicable for intense networks and also effective for encrypted communication[14, 15].

Data Mining for Intrusion and Malicious Code Detection

A number of tools have already been developed that use data mining for cyber security applications at the University of Texas at Dallas [3], 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. 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 tool examine emails and extracts 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 it can be tested to determine whether the email has a virus/worm. Tool uses training and testing data sets posted on various web sites [13]. For firewall policy rule analysis tool use association rule mining techniques to determine whether there are any anomalies in the policy rule set [1].

Existing Systems

In this section, we present some of the implemented systems that apply data mining techniques in the field of Intrusion Detection.

- **ISOA (Information Security Officer's Assistant)** [22]: ISOA is a system for monitoring security relevant behavior in computer networks. ISOA serves as the central point for real-time collection and analysis of audit information. When an anomalous situation is identified, associated indicators are triggered. ISOA automates analysis of audit trails, allowing indications and warnings of security threats to be generated in a timely manner so that threats can be countered. ISOA allows a single designated workstation to perform automated security monitoring, analysis and warning
- **Distributed Intrusion Detection System (DIDS)** [18]: A risk intrusion detection system that aggregates audit reports from a collection of hosts on a single network. Unique to DIDS is its ability to track a user as he establishes connections across the network.
- **EMERALD** [16]: EMERALD is a software-based solution that utilizes lightweight sensors distributed over a network or series of networks for real-time detection of anomalous or suspicious activity. EMERALD sensors monitor activity both on host servers and network traffic streams, and empower system defenders with the capacity to detect and ultimately thwart cyber attacks across large networks By using highly distributed surveillance and response monitors, EMERALD provides a wide range of information security coverage, real-time monitoring and response, protection of informational assets. EMERALD implements an enterprise-wide analysis to correlate the activity reports produced across asset of monitored domains. EMERALD offers protection from network-wide threats such as Internet worm-like attacks, attacks repeated against common network services across domains, or coordinated attacks from multiple domains against a single domain.
- **The MINDS System** [10]: The Minnesota Intrusion Detection System (MINDS), uses data mining techniques to automatically detect attacks against computer networks and systems. While the long-term objective of MINDS is to address all aspects of intrusion detection, the system currently focuses on two specific issues:
  – An unsupervised anomaly detection technique that assigns a score to each

network connection that reflects how anomalous the connection is, and

- An association pattern analysis that summarizes those network connections that are ranked highly anomalous by the anomaly detection module.

Experimental results on live network traffic at the University of Minnesota show that the applied anomaly detection techniques are very promising and are successful in automatically detecting several novel intrusions that could not be identified using popular signature-based tools such as SNORT. Furthermore, given the very high volume of connections observed per unit time, association pattern based summarization of novel attacks is quite useful in enabling a security analyst to understand and characterize emerging threats.

- **The IDDM project [20]:** The IDDM (Intrusion Detection using Data Mining) project is a project that uses data mining techniques in order to describe the data on a network and analyze them for further deviation in observed traffic. IDDM utilizes meta-mining to achieve its goals. The goal is to track and understand changes in the network traffic over time. IDDM uses association rules in order to observe network traffic. Two different snapshots of the association rules -created in two different timestamps- are compared in order to see which rules have remained the same, have been changed, been added and which have been eliminated. The system uses agents that apply association rule mining on raw network packets. Attributes that are taken into consideration are:

  - Packet type (protocol)
  - Source/Destination Port
  - Packet Size
  - TCP flags

  Analysis on a stable network should produce the following results between the 2 snapshots:

  - A small number of added/deleted rules
  - A Fairly large number of unchanged rules
  - A Small to medium number of changed rules

  Two more systems that use the basic meta-mining concept behind IDDM can be found at :

  - http://www1.cs.columbia.edu/jam/itoprojsubmitted99.html
  - The MADAM ID System is part of the larger JAM project. It is held at the department of Computer Science at the University of Columbia and is led by Prof. Salvatolre Stolfo [19].
  - In Iowa State University researchers Helmer et al [5] use agents in order to collect low level information and correlate them at a higher level.

- **IDSs in the Open Market:** Various systems that employ data mining techniques have already been released as parts of commercial security packages. Some of the most popular of these systems are:

  - RealSecure SiteProtector [6]
  - Symantec ManHunt [7]
  - nSecure nPatrol,
  - Dshield [8]
  - MyNetWatchman [9]

Conclusions

This paper has discussed data mining, knowledge discovery and tool that can be used in cyber security and network intrusion detection. Paper starts with brief introduction to data mining and knowledge discovery and then puts light on various cyber security issues. Brief overview has been given about various data mining tools that have been designed to detect cyber security intrusion.

References

1. 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.

2. Agrawal, R.; Mannila, H.; Srikant, R.; Toivonen, H.; and Verkamo, I. 1 996. Fast Discovery of Association Rules. In *Advances in Knowledge Discovery and Data Mining,* eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 307–328. Menlo Park,Calif.: AAAI Press.

3. Bhavani Thuraisingham, Latifur Khan, Mohammad M. Masud, Kevin W. Hamlen “Data Mining for Security Applications” IEEE/IFIP International Conference on Embedded and Ubiquitous Computing2008

4. Cios, K.J., Teresinska, A., Konieczna, S., Potocka, J., Sharma, S., Diagnosing Myocardial Perfusion from PECT Bull’s-eye Maps - A Knowledge Discovery Approach, IEEE Engineering in Medicine and Biology Magazine, Special issue on Medical Data Mining and Knowledge Discovery, 19:4, pp. 17-25, 2000

5. Helmer G.G, Wong.J.S.K, Honavar V, Miller L., Intelligent Agents for Intrusion Detection”, Proceedings of the IEEE Information Technology Conference, Syracuse, USA.

6. http://www.afina.com.ve/download/docs/iss/iss real%20secure.pdf

7. http://www.softwarespectrum.com/business/TAAP Library/Symantec docs/Manhunt Fact Sheet.pdf

8. http://www.dshield.org/

9. http://www.mynetwatchman.com/

10. Levent Ertoz and Eric Eilertson and Aleksandar Lazarevic and Pang-Ning Tan and Vipin Kumar and Jaideep Srivastava and Paul Dokas, “MINDS- Minnesota Intrusion Detection System”, Next Generation Data Mining, MIT Press, 2004.

11. Liu Wenjun: “An Security Model: Data Mining and Intrusion Detection”,2nd International Conference on Industrial and Information Systems .PP. 448-450, July 2010

12. 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.

13. 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.

14. Masud, Mohammad, Al-khateeb, T., & Khan, L. (2008). Flow-based identification of botnet traffic by mining multiple log files. , 2008. DFmA 2008., 200-206.

15. Masud, MM, Gao, J., Khan, L., & Han, J. (2008). Peer to peer botnet detection for cyber-security: a data mining approach. CSIIRW 08.

16. Porras, A. and Neumann, P. G., “EMERALD: Event Monitoring Enabling Responses to Anomalous Live Disturbances”, In Proceedings of the National Information Systems Security Conference, October 1997.

17. Schwartz, Mathew J. "Microsoft Leads Zeus Botnet Server Shutdown." Information Week. N.p., 26 Mar. 2012. Web. 9 Nov. 2012.

18. Snapp, S. R., Smaha, S. E., Grance, T., Teal, D. M., "The DIDS (Distributed Intrusion Detection System) Prototype", In Proceedings of the USENIX Summer 1992 Technical Conference, pages 227-233, June 1992.

19. S. Stolfo, A. L. Prodromidis and P. K. Chan, "JAM: Java Agents for Meta-Learning over Distributed Databases", in Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, D. Heckerman, H. Mannila, D. Pregibon, and R. Uthurusamy, editors, AAAI Press, Menlo Park, 1997.

20. Tamas Abraham, "IDDM: Intrusion Detection using Data Mining techniques", Information Technology Division Electronics and Surveillance Research Laboratory, May, 2001.

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

22. Winkler, J. R., Landry, L. C., "Intrusion and anomaly detection, ISOA update", In Proceedings of the 15th National Computer Security Conference, pages 272-281, Oct. 1992.