Intelligent Strategies for 21st Century Broadband and Cyber Infrastructure Security

Dr. Emmanuel Hooper
Senior Scholar and Researcher
Cyber, Broadband and Intelligent Security
Harvard University, Leadership for Networked World
Harvard-MIT-Yale Cyber Scholar
Founder, Global Information Intelligence
http://globalinfointel.com
ehooper@fas.harvard.edu or ehooper@aya.yale.edu

1 Introduction

1.1 The Challenges Facing Broadband and Cyber Networks

This research is pertinent to recent 21st cybersecurity challenges and the FCC broadband plans to gather data-driven information on how the public safety community can best utilize broadband technologies towards preparation, response, and recovery from major disasters, pandemics, acts of terrorism, and cyber attack. This includes challenges of broadband and cyber security for the FCC, Cyber Coordination Executive, National Cyber Study Group (NCSG), Director of National Intelligence (DNI), and Comprehensive Cybersecurity Initiative (CNCI). The acceleration of high speed broadband networks using wireless, Wi-Fi, and emerging WiMAX and their interconnection to cyber infrastructures present many critical challenges. It is vital that in the proposed Broadband plan being presented to Congress by the FCC that continuous research and study of the impact of relevant issues are discussed. These include the challenges of high speed data transmission at high frequencies and speeds in the Gigabits range that transfer eventually terabytes of data across multiple interfaces to wireless and fiber networks that interconnect to the global Internet. In the midst of the emerging challenges for 21st century US Cyber Security, broadband data transmission and its intersection critical US infrastructures present many issues that require detailed on-going research. This includes broadband distribution, access controls, security, monitoring, intrusion detection and prevention, response and forensic evidence for traceability, sustainability of effective use and management. Among the many challenges ensuring responsible broadband use, distribution and management is monitoring astute interceptions of high-speed traffic at various segments of the broadband infrastructures that interface with US cyber and global networks that transmit high-speed data in real time. The challenges here include identification of legitimate traffic, understanding the levels of appropriate thresholds for traffic on broadband networks, and ensuring effective cryptographic key management, ciphers and algorithms are adaptable to handle astute interceptions, evasions and insertions. Additional challenges include congested networks and packets in backbones of cyber networks due to the rapid increase in the number of users, types of services, multi-functional applications, and meta data aggregation servers in regional and global data centers. Furthermore, coupled

with the increasing trends of parsing, loading and transforming government, corporate, public and private data at high speeds between laptops, PDUs and Data Centers, the interception of high speed packets can easily evade detection and appropriate response in real-time.

1.2 Emerging Challenges for Broadband and Cyber Security

While availability and rapid dissemination of high-speed digital networks has been a major focus of consumers and other advocates, the lessons of attacks on cyber networks, and over 5,000 categories and subcategories of attacks should cause 21-century policy makers prioritize on effective broadband traffic distribution and security. Security software and hardware have increasing vulnerabilities due to the lack of strategic design for adaptable performance for threat mitigation and ineffective anti-hacking code security standards. Coupled with the multi-functionality, security applications and lack intelligent data mining to detect and respond to astute hackers who operate in stealth mode using broadband networks to penetrate Virtual Private Networks (VPNs), dedicated global backbones and cyber infrastructures. The major challenges for broadband use, distribution, dissemination and integration with emerging cyber infrastructures include the vulnerabilities in software, hardware, operation systems, applications, databases, routers, firewalls, Intrusion Detection Systems (IDS), Intrusion Prevention Systems (IPS), Intrusion Response Systems (IRS), log aggregation mechanisms, event correlation, parsing and correlating rules, rule induction, hybrid algorithms, intelligent data mining, forensic evidence data acquisition and analysis of broadband high speed data. The attacks are well beyond Distributed Denial of Service (DDOS), buffer overflows, cross-site scripting or cross-reference scripting in application security for broadband traffic. Specially designed attacks using embedded stealth software and malware, and crafted by rogue interceptions of broadband signals crossing backbones of cyber traffic has challenged and evaded many Internet Service Providers (ISP). Additional challenges facing broadband traffic includes the emerging trends in data aggregation, using security event and incident management software and Meta or Master Data Management (MDM) at regional, national and global centers. This implies that very sensitive data are transmitted at high speeds into centralized storage. There are several problems posed by such trends. Since broadband traffic is traveling at high speeds, real-time the detection and response to attacks versus false positives, or normal traffic versus abnormal traffic can be ineffective. This implies that increasing reliance on automated software, which largely depend on rudimentary rules rather than hybrid adaptable algorithms, make the task of broadband traffic management both inefficient and ineffective. Furthermore, increasing challenges include ”man-in-the-middle” interceptions at interfaces of broadband traffic across wireless, Wi-Fi, WiMAX, and autonomous backbones that interface with secured interfaces. Even passphrases of 128 characters, and encrypted data can be readily decrypted using forensic tools using astute techniques.

1.3 Current Approaches and Their Problems

Current methods for detecting broadband attacks and malicious activities lack effective collaborative responses [3, 15, 32, 28], due partly to increase in volumes of normal traffic [15, 32, 36]. This includes exponential increase in packet rates from astute attacks on fast broadband networks, VPNs and Internet and autonomous backbones [8, 27, 32, 34, 37].

2 New Intelligent Strategies for 21st Century Broadband and Cyber Security

Intelligent and strategically new approaches are required to meet the rapidly emerging challenges and threats for broadband technology, traffic, networks and cyber security in the 21st century. The multi-functional features in emerging software and hardware designed for broadband users, also offer the ca-

pability for astute users to utilize these functions to spy, evade detection, engineer stealth attacks and even prevent traceability in real-time. Most of the real-time event analyzers are incapable of tracking astute real-time attacks due to increasingly aggregated data without effective hybrid algorithms consisting of adaptable pattern recognition and matching for handling astute event anomalies. Thus in broadband speeds, short-term attacks of insertions and intermittent stealth evasions are readily concealed, since the gigabytes of exchanged traffic at high speeds are aggregated to terabytes stored data, lacking effective hybrid data mining techniques to filter out feature attributes for responding to astute broadband threats in real-time. This implies strategic intelligent algorithms, for access controls, dissemination, adaptability and rapid response to broadband cyber threats and counterintelligence. Research indicates that the amount of traffic exceeds terabytes per second at many ISPs, data centers, corporate and public networks that use broadband via wireless, Wi-Fi, or fiber networks to interconnect to their local, regional, national and global networks. Furthermore, the transmission of broadband traffic over secured WiMAX, present challenges of internal traffic security. Astute hackers evade access control systems using patterns of infrequent attacks to conceal penetration into centralized data at traffic speeds of Gigabits/second. At such speeds an IDS, IPS, IRS, Security and Event Management Systems, Log Aggregation and Correlation Systems, are vulnerable to false positives and false negatives.

1. **Effective detection and monitoring techniques for US Cyber transfers of highly sensitive data**: First, embedded monitoring and filtering detection mechanisms, techniques, countermeasures and counter-intelligence strategies against evasive interception of highly sensitive data, consisting of private, confidential and personally identifiable information for customers, consumers and products at intermediary points of US Cyber critical infrastructures.

2. **Specification of new privacy and controls for secure data transmission**: Secondly, specification of new privacy definitions and control requirements in categories, subcategories and attribute types for secure transmission against interception of highly sensitive data and to ensure that information flow of highly sensitive data in complex systems comply with newly derived security and privacy policies and procedures.

3. **Validation and verification techniques for effective privacy controls for secure data transmission**: Thirdly, techniques for validation and verification of effective implementation of these privacy specification controls based on new monitoring techniques for enforcement of security and privacy controls transparency, auditability, traceability accountability and traceability and cyber forensics and breaches assessments and notifications and preventions in critical infrastructures, applications and databases.

These implies the design of new security control mechanisms that specifies new security, privacy includes new terms, need-to-validate" and need to-verify adequate effective controls and validation evidence for security and privacy of highly sensitive data across broadband and cyber critical infrastructures.

2.1 Intellectual Merits

The intellectual merit of the research includes contributions to research and development of effective solutions to the major emerging and rapidly increasing problems of intelligent broadband and cyber attacks, most of which evade detection in complex infrastructures carrying data-intensive applications. This research will enhance understanding, analysis and effective design of complex infrastructures operating in the context of cyber security and protection design. Effective approaches for secure transmission of highly sensitive data, carrying private, confidential and personally identifiable and customer

and product class. This includes privacy controls and trustworthy transaction processing, security policy specification, discovery, and implementation, and generation of automatic of security configurations and responses from abstract security policies and control specifications for international data transfer controls implementation, and detection and responses to counter measures against interception of highly sensitive data, carrying private, confidential and personally identifiable and customer and product class information at intermediary points. This includes effective cybersecurity and protection design, and intelligent strategies for complex network infrastructures that transfer very sensitive personally identifiable information.

2.2 Broader Impacts

The broader impacts resulting from the research include effective 21st century solutions for US broadband and cyber security, FCC regulations, DHS, DNI, DOD,DARPA, IARPA, DHSARPA, etc. This involves effective solutions to the challenging and rapidly increasing problems of evasive and intelligent cyber attacks that face cyber security and protection design, complex network infrastructures which carry data-intensive applications across the Internet. Furthermore, the broader merit of the research results will be efficient, resilient, robust, scalable, adaptable, accurate and provide effective intelligent counter-attack solutions, involving both real-time responses and off-line for forensic analysis for current and future research on the design and effective implementation of broadband and complex network infrastructure security and data-intensive applications across the internet. This includes discovery, understanding teaching, training and learning and broadens participation of underrepresented groups including various gender, ethnicity and geographic regions. Benefits to Society includes solutions of the major problems of cybersecurity, cyber-crime, counter-terrorism, identity theft, US Cyber data transfer of data on the Internet, industrial data and applications and education. This includes Partnership with academic scientists, staff at federal agencies and with the private sector on both technological and scientific national projects and interpretation of research and education results in formats understandable and useful for non-scientists and provide information for policy makers and formulation by International, federal, State and local agencies on regulations on Internet security, cyber crime and countermeasures.

3 Analysis of Previous Work

The approach for solving the problem of benign and suspect network traffic [10] has focused on attacks at the perimeter of the network, which is the boundary between the external routers and the Demilitarized zones (DMZ). The DMZ consists of hosts, which provide various services designed for public access (hosts from external network). The DMZ is separated from the internal network via firewalls and security controls in order to restrict public access to the internal network. These perimeters receive significant attacks of Distributed Denial of Service (DDOS) attacks, including DDOS passwords attacks, DDOS brute force and malware, and astute attacks. These are aimed at bypassing firewalls, IDSs and VPN interface routers and gateways to internal VPN servers in order to access highly sensitive data. Attacks and interception and eavesdropping of highly sensitive data occur at perimeters and also at internal interfaces to US Cyber routers, and backbones of external networks and the Internet.

3.1 Limitations of Detection Systems, Tools and Techniques

Major problems in various IDSs and intrusion detection tools include generating false positives and false negatives, accumulating benign traffic and low detection accuracies, especially in broadband network infrastructures environments. Moreover, IDSs are vulnerabilities to hackers and inefficient in their handling of benign traffic and responses to attacks [20, 37, 45]. In addition, various IDSs, using largely

signature detection mechanisms, face the problems of accurate detection in real-time and presenting appropriate responses for containment and mitigation of broadband network infrastructure attacks. For example, IDSs which handle large data sizes, ranging between several Megabits to several Gigabytes per second, also drop packets and decrease in detection accuracy as packet rate increase [20]. This includes network IDSs such as Intrushield IDS [32, 32], Realsecure [30], NetRanger [4], Blackice [29], snort [39], Bro [34] and NFR [33, 38], Graph-based IDS (GR-IDS) [43], NetSTAT [48], Quicksand [23, 24] and SPARTA [23].

3.2 Distinguishing between malicious and random benign broadband traffic

Other IDS tools are incapable of clearly distinguishing between malicious connections and random but normal variations in broadband traffic patterns. Thus, they are subject to significant false positives and false negatives. This includes host IDSs such as DIDS (Distributed IDS) [41], Intrusion Detection Agent (IDA) [1], SWATCH [12], COAST [25] and USTAT [19]. Furthermore, detection systems that aim to respond to alerts do not adequately determine the real status of the alerts before attempting to provide any response. Thus they do not provide effective responses based on corresponding rigorous traffic analysis. Thus they mistakenly respond to alerts that have significant false positives. This includes NAI [32, 31], Realsecure [30], NetRanger [4], Blackice [29], NFR [33, 38], Netstat [48], STAT [47] and SWATCH [12]. Similarly, IDS tools that attempt to provide some responses have limitations due to the specificity or over-generalization of their model of the signature patterns. Thus, they are unable to determine the real status of a wide variety of alerts. This includes Norton Internet Security [44], Intrusion Detection Agent (IDA) [1], Intrusion Detection Alert [35], CITRA [9], IDIOT [26], INSA [46] and Ji Nao [21]. In summary, none of these detection tools and techniques provides feedback mechanisms along with rigorous statistical analysis. Analysis of detection strategies for various traffic types indicate that both anomaly and misuse detection are relevant.

3.3 Analysis of strategies for detection and response to different broadband attacks

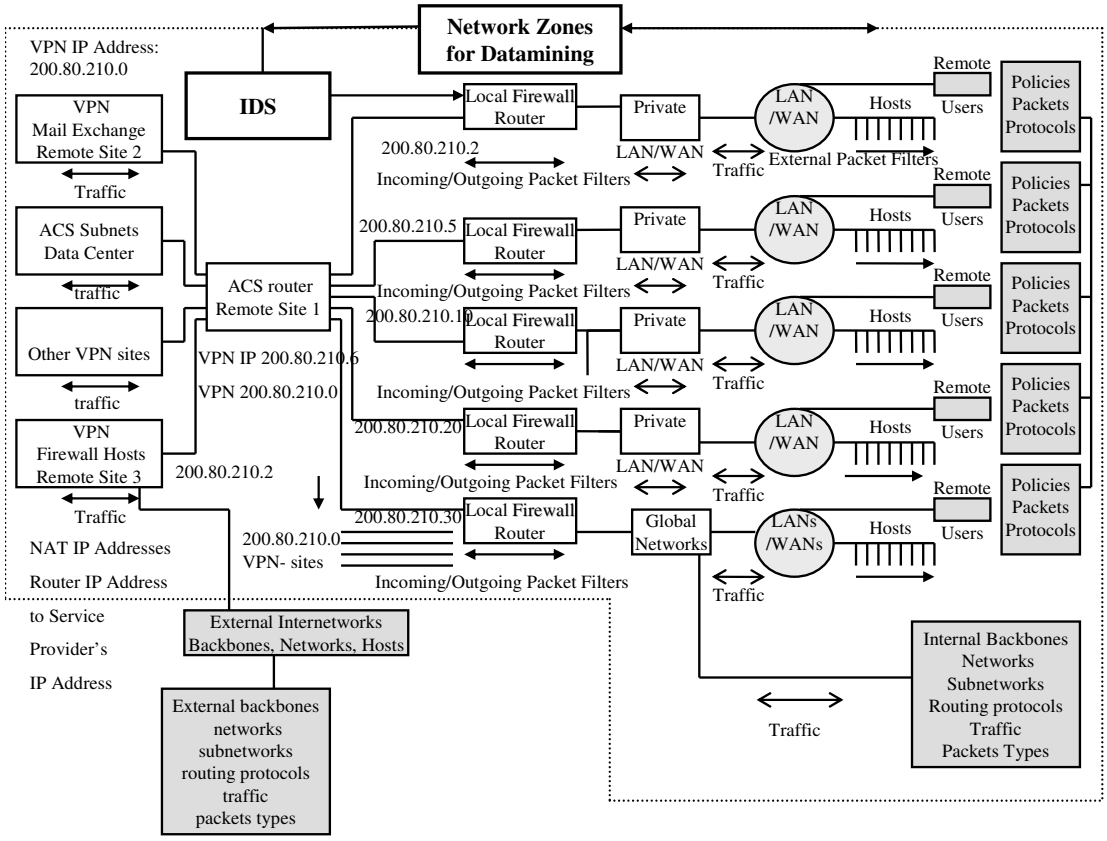
In order to generate appropriate responses to suspect packets at the perimeter, various strategies are required for analysis and detection of different broadband attack types. Appropriate responses should correspond to the various packet types whose attributes require analysis, prior to their development and implementation. Attack types are now more complex due to the proliferation of applications and network systems and their associated vulnerabilities [10, 22].

4 Architecture

4.1 Highly Sensitive Broadband Traffic Data Mining on Autonomous Network Interfaces and the Internet

This involves strategic traffic interception, filtering and monitoring of broadband traffic on wireless, Wi-Fi, WiMAX, Autonomous Networks and Internet Interfaces routers, VPNs and backbones. The strategic response approach involves the prototype design for monitoring and filtering traffic for specific attribute features of packets and malicious traffic from VPNs, firewalls and IDSs. The packets are diverted for analysis in the data mining analysis database hosts. An example of strategic architecture for monitoring, filtering and data mining of highly sensitive information from multiple VPNs, firewalls and IDS is shown in Figure 1.

Architecture for Highly Sensitive Information Datamining:
Strategic Traffic Interception, Filtering and Monitoring on
VPN, Autonomous Networks and Internet Interfaces



6nsf.pdf

Figure 1. Data Mining of Highly Sensitive Data: Multiple VPNs, Firewall, IDS and Autonomous Network Interfaces

This shows multiple firewall zones for local and remote sites within the VPN carrying broadband traffic. The external traffic is separated from the internal network through the Private Internet Exchange (PIX) firewall [6]. The firewall is placed in internal and external switch interfaces in the DMZ to isolate the external traffic from the internal traffic. This includes multiple firewalls for local and remote sites within the Virtual Private Network (VPN). The local and remote routers isolate traffic between the LANs, MANs and WANs. The remote firewalls track access violations within these zones, while the local firewalls trace and respond to attacks in the DMZ and the internal network.

5 Research Methods

The research method consists of developing new and effective embedded monitoring and filtering detection mechanisms, techniques and countermeasures against evasive interception of highly sensitive data, consisting of private, confidential and personally identifiable information for customers, consumers and products at intermediary points of US Cyber critical infrastructures. Secondly, the specification of new privacy definitions and control requirements in categories, subcategories and attribute types for secure transmission against interception of highly sensitive data and to ensure that information flow of highly sensitive data in complex systems comply with newly derived security and privacy policies and procedures. Thirdly, the development of new and effective techniques for validation and verification of effective implementation of these privacy specification controls based on new monitoring techniques for enforcement of security and privacy controls transparency, auditability, traceability accountability and traceability and cyber forensics and breaches assessments and notifications and preventions in critical infrastructures, applications and databases.

5.1 Data Mining of Broadband Traffic in Autonomous and Cyber Networks

The autonomous networks route packets from broadband traffic and multiple backbone networks via the Internet. However, these autonomous broadband networks also are vulnerable to attacks from hackers. The multiple firewalls, subnets strategically isolated to divert traffic into different zones for analysis. The additional backbones for the US Cyber networks supply further traffic to the VPN. Thus the additional traffic between the local site VPN, the remote VPN site and the remote Private LAN/WAN sites is filtered using the "local firewall router" at each site. The cumulative traffic is diverted to the Data Mining Database Hosts (DDH) for analysis and the results are sent to the IDS for responses based on their statuses. This ensures that any hackers attempting to penetrate the local VPN using the remote sites are denied access to the VPN. The additional remote traffic is controlled, segregated and filtered using Cisco Secure Access Control Servers (ACS) [7] routers to prevent access to sensitive segments of the internal VPNs.

5.2 Highly Sensitive Information Broadband Traffic Interception, Filtering and Data Mining

This involves using multiple firewalls for local and remote sites within the Virtual Private Network (VPN) to intercept, filter and perform hybrid data mining on broadband traffic. These critical segments of network infrastructures include the VPN and the local sites, private networks within the VPN, including local and remote sites in the VPN that share private data subnets, and thus unique private networks within the VPN. Each subnet is designed with a range of IP addresses, corresponding subnet mask and default gateway. These separate traffic for efficient log management and traceability of the packet as it traverses multiple zones between source and destinations. It also prevents the packet from loading the network and provides means for its isolation and analysis using packet filters for each zone. The packet from each zone

is diverted to the DDHs for analysis. Remote traffic to the VPN using the Cisco Access Server (AS5301) via Remote Authentication Dial In User Service (RADIUS) [5] is diverted to the DDHs for analysis of potential malicious traffic. This prevents malicious packets from using backdoor access to the internal network through dial-up access or remote VPN firewall access. The traffic in the various zones and VPN zones is analyzed through the DDHs for potential malicious packets. In order to provide efficient management of the traffic volume, the patterns of the traffic are examined for similar packets using the data mining techniques. These are used to filter similar patterns of traffic which are subsequently diverted to the DDHs. This implies that the DDHs focus on examining new malicious traffic and the IDS filters out any previously known patterns from its alert monitor. Thus repetitive well-known benign traffic is filtered and permitted access to their destinations and known attacks are diverted to the IDS for denial of access to their intended destinations. This strategy reduces traffic in the DDH hosts and enables efficient management of packets entering or leaving the multiple zones of the network infrastructure.

6 Intelligent Broadband Infrastructure Protection Strategy for Highly Sensitive Data

We analyzed real broadband network traffic in a commercial environment consisting of Intrushield IDS [32]. The experiment consists of alert data collection and analysis in a network environment comprising the following architecture (see Figure 2). The IDS consists of Network Associates Intrushield (NAI), "1-2600 IDS" [32] for detecting attacks from external traffic, located at the Demilitarized Zone (DMZ), "IDS-DMZ", and two main internal IDSs (NAI-I-4000 IDS), which detect internal attacks (IDS-4001 and 4002). The "1-2600 IDS" sensors monitor traffic at 600 megabits per second and the I-4000 IDS sensors monitor traffic at 2 gigabits per second [32]. The alerts of suspicious traffic are monitored and analyzed in the IDS management consoles: the Main IDS Management console (Mgt-01) and the backup IDS console (Mgt-02). These alerts are stored in the IDS database of the DDH Attack Pattern Database for further analysis. The internal Virtual Private Network (VPN), Virtual Local Area Networks (VLANs) and subnets are separated from external traffic by the DMZ Firewalls – Private Internet Exchange (PIX) firewalls [6] and internal firewalls – Checkpoint FW-1 (CP Firewall); proxy servers, Internet Security Acceleration (ISA) servers), routers, load balancers, gateways and switches at the perimeter. The IDS logged over 100 gigabytes of 60 days of aggregated logs from the firewalls, IDSs, routers, load balancers, gateways, switches, and application and data server logs. These were analyzed for benign traffic by comparing them with the normal long-term characteristics in that network environment. The alerts were analyzed for benign traffic by comparing them with the normal long-term characteristics in that network environment. Analysis of the characteristics of the network traffic indicated a high number of false positives because much of the suspect traffic was actually from benign hosts.

7 Steps in Experimental Methods: Data Mining Analysis including Classification, Clustering, Genetic Algorithm, Pattern Generation and Analysis, Rule Induction and Statistical Analysis

Pre-processing and Formatting the Datasets: Selection and Sorting of Attack/Intrusion or Normal Data

Data Selection: initially, about 10% of 100 Gigabyte of the data from commercial environment of VPNs interfaces to broadband networks was selected at random for analysis. From this dataset, 40% was selected at random as training set and the remaining 60% was used as the test set. None of the cases in the training set was used for testing and vice-versa. Specific training cases were selected at random and exported into SPSS software [42] and Rosetta toolset [18] for analysis.

7.1 Subcategories and Distinctive Attacks Attributes

The resultant summaries of the statistical analysis, correlations and commonly occurring attributes for each subcategory. These distinctive features of attributes for the various subcategories were analyzed to detect unknown attacks. A program was written to examine the efficiency and accuracies of the subcategories.

7.2 Analysis of attack types, categories and subcategories

Suspect packets, both benign and malicious [2], are identified in terms of potential intrusion/attack types, categories and subcategories [32]. Since connections may be abnormal, the traffic from the source host that exhibits characteristics similar to these attacks can be isolated from the broadband network to the quarantine channels for analysis. Different packet types require different responses, such as divergence to hosts for Denial of Service (DOS) attacks [11] or responses from quarantine zone hosts with specific responses for buffer overflow attacks.

7.2.1 Attack types and categories

The most current attack/intrusion types are as follows: statistical-anomaly, application-anomaly, protocol-anomaly, multi-method-correlation, multi-flow-correlation, signature anomaly and threshold anomaly. See Table 1 below for their definitions. The corresponding attack categories for the most current attack/intrusion types are as follows: exploit, policy violation, reconnaissance and volume DOS and DDOS. The suspicious attack types and categories are more accurately defined through their subcategories. Since there are over 10,000 attacks with similar attributes [32], analysing their subcategories is more effective for rapid identification and appropriate response.

Table 1. Attack types

Attack Detection Types	Description	Example
Statistical-anomaly	A change in percentage composition of traffic flow – statistical changes, e.g., Volume DOS attack.	Volume DOS Attacks
Application-anomaly	Differences in applications codes, anomalies	SMTP: Too Many Commands
Protocol-anomaly	Abnormal protocols, contents headers and data segments	RADIUS Buffer Overflow
Multi-method-correlation	Multiple methods in correlated attacks	RLOGIN buffer overflows
Multi-flow-correlation	Multiple packets in correlated attacks	ICMP Host Sweep
Signature	Attacks identified by string patterns	BACKDOOR Windows Shell exploit
Threshold	Attacks exceeding typical threshold values	TCP Port Scan

8 intelligent infrastructure protection strategy for highly sensitive data on Broadband networks

We analyze real broadband network traffic in a commercial environment consisting of Intrushield IDS [32]. The experiment consists of alert data collection and analysis in a network environment comprising

the following architecture (see Figure 2). The IDS consists of Network Associates Intrushield (NAI), "1-2600 IDS" [32] for detecting attacks from external traffic, located at the Demilitarized Zone (DMZ), "IDS-DMZ", and two main internal IDSs (NAI-I-4000 IDS), which detect internal attacks (IDS-4001 and 4002). The "1-2600 IDS" sensors monitor traffic at 600 megabits per second and the I-4000 IDS sensors monitor traffic at 2 gigabits per second [32]. The alerts of suspicious traffic are monitored and analyzed in the IDS management consoles: the Main IDS Management console (Mgt-01) and the backup IDS console (Mgt-02). These alerts are stored in the IDS database of the DDH Attack Pattern Database for further analysis. The internal Virtual Private Network (VPN), Virtual Local Area Networks (VLANs) and subnets are separated from external traffic by the DMZ Firewalls – Private Internet Exchange (PIX) firewalls [6] and internal firewalls – Checkpoint FW-1 (CP Firewall); proxy servers, Internet Security Acceleration (ISA) servers), routers, load balancers, gateways and switches at the perimeter. The IDS logged over 100 gigabytes of alerts in 60 days.

8.1 Applying Data Mining Techniques

Data Mining Techniques: The Data Mining Analysis involving Classification and Genetic Algorithm [18, 49] in Pattern Analysis for known attacks. Rule Induction using C5.0 [40], K-means clustering [13, 14] and the Genetic Algorithm for computing a fitness function [18, 49] were used for detection and analysis of categories and subcategories of anomaly patterns. For detection of subtle and complex attacks a Framework of Hybrid consisting of Rule Induction using Holte's 1R rule [16] and Statistical Analysis [17] were applied via the Rosetta toolset [18], followed by filtering for maximum support of conditional attributes to increase accuracies. Various cases of attack and normal classes were selected at random and algorithms were applied to each class type. This produces a set of decision rules or general patterns via minimal attribute subsets that distinguish on a per object basis. This is followed by filtering rules with maximum support for each class in order to obtain an optimum set of conditions for each ruleset for an attack/normal class. This was followed by development of matrix - table of conditions for each attribute in rulesets. Subsequently, for each attribute value if then rules were developed based on the attribute values each conditional ruleset. A program was written using the conditional rules from the Table (Matrix) of rulesets for each class of the specfied cases in the training data. Finally, there was validation of the accuracies of detection of class types using test data. See summary of Results in Table 2.

9 Preliminary Research and Results

The preliminary research results for detection accuracies using various data mining techniques are shown in Table 2.

The difference between a statistical anomaly and an exceeded threshold is that the statistical anomaly results in the pattern of traffic being abnormal to the normal patterns or values, such as the rate of normal traffic for TCP or UDP sessions directed to specific interfaces or IP addresses and subnets in a particular amount of time. This consists of multiple statistical factors deviating from the normal expected pattern. On the other hand, an exceeded threshold refers to a specifically pre-defined volume or quantity, such as the limit for acceptable levels of protocols that can be considered tolerable in a network. An example is a pre-defined amount of ICMP packets per second that can be defined for a network.

9.1 Statistical Techniques for Detecting Complex and Unknown Attacks

The statistical techniques in this experiment involve statistical cross-correlation and analysis of relevant and critical attribute features in categories and subcategories of attack types. This was followed by statistical isolation of distinctive attribute features and pattern analysis of various attack subcategories.

Table 2. Results: Data Mining Techniques Accuracies Summary

Data Mining Technique	Number of Cases	Training Accuracies	Number of Cases	Testing Accuracy
K-means Clustering	34,035	**94.83%**	47,050	**92.74%**
Cubits/C5.0 Rule Induction	40,470	**92.69%**	79,013	**90.58%**
1R Rule Induction	45,471	**91.78%**	79,618	**90.21%**
Hybrid 1R Rule Induction and Filtering for Maximum Support	2,422,535	**99.90%**	3,334,037	**99.89%**
Application of Hybrid to Real IDS Datasets-Application Security Features	11,289	**97.97%**	21,423	**99.88%**
Application of Hybrid to Real Firewall Datasets- Network Security Features	10,746	**100.00%**	14,392	**99.97%**
Correlation of Real IDS Datasets and Firewalls Security Features	18,465	**99.16%**	29,700	**98.51%**
Application of Hybrid to Real TCP Attack Flags Security Protocol Features	11,650	**100.00%**	21,500	**100.00%**
Application of Hybrid to Real Intruders Network Patterns	2,422,535	**99.90%**	3,334,037	**99.89%**
Application of Hybrid Correlation of Statistical Anomaly and Signature Detection Analysis	80.400	**99.73%**	153,515	**99.57%**

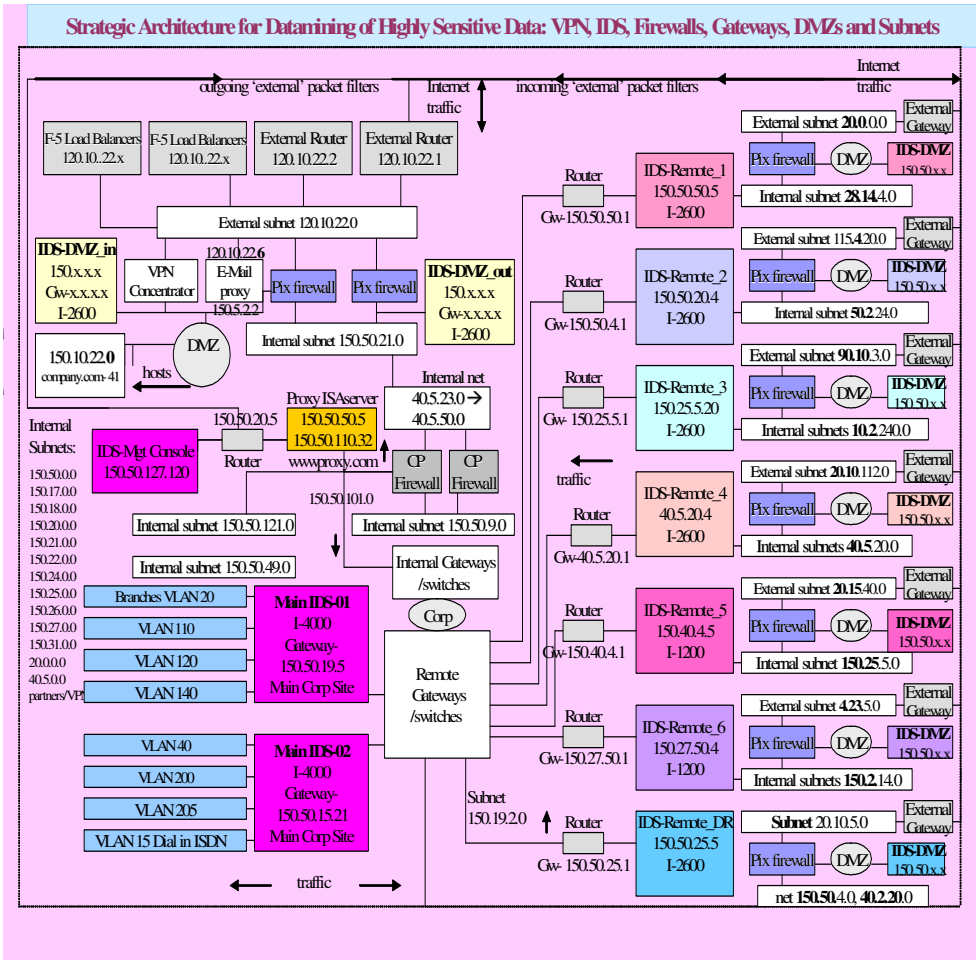
The resulting model from the training sample of frequencies and occurrences of unique characteristics for subcategories of attacks used to identify unknown attacks in the test sample. The statistical correlation of categories and detection mechanism indicates that highest category is policy violation which consists largely of the signature detection mechanism, followed by protocol anomaly. See Figure 2. Although multi-flow correlation- reconnaissance is second highest to signature-policy violation attacks, overall, the exploits category exceeds the reconnaissance category, which comprises application and protocol anomalies and signature attacks. This is because exploits are major attacks while reconnaissance are preliminary probings. Volume DOS, comprising statistical anomalies, is the lowest attack category, since its effect is mostly disruption of service rather than gaining access to private data.

10 Research Agenda

The general plan of work of this robust research is on critical US Cyber infrastructures protection involving the development of new approaches for secure transmission of highly sensitive data, carrying private, confidential and personally identifiable and customer and product class. The plan of work includes research, analysis, monitoring, evaluation, and reporting and development of effective security and privacy controls for transferring highly sensitive, private confidential customer and product involving secure transmission and verification US Cyber traffic as follows:

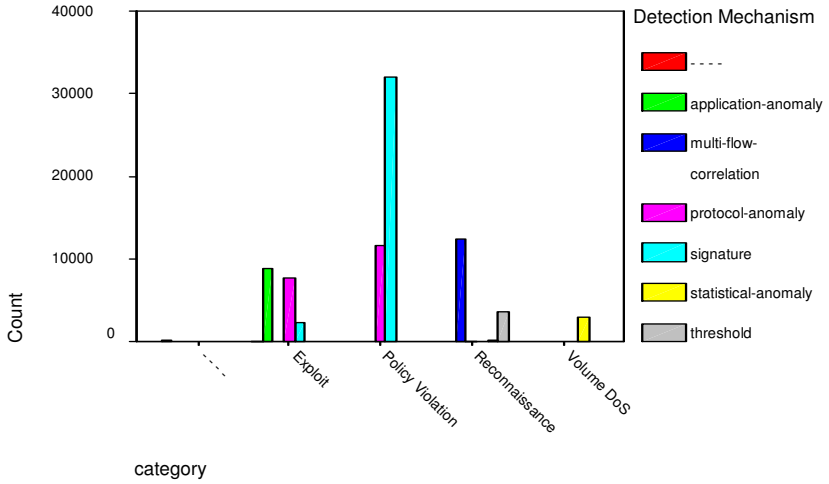
1. Development of new Cyber privacy controls transmission of highly sensitive data including trustworthy transaction processing, security policy specification, and discovery and implementation validation.

2. Development of embedded secure cyber monitoring, detection and response software filtering for counter measures against interception of highly sensitive data, carrying private, confidential and



Statistical Data Mining: Correlation of Categories and Detection Mechanism

The "Count" field is the frequency of attacks in number of connections



Data Mining.pdf

Figure 2. Efficient Data Mining Analysis

personally identifiable and customer and product class information at intermediary points .

3. Filtering of interesting cyber packets, using correlation and aggregation of analysis of interception of autonomous backbone traffic.
4. Analysis and validation and verification of secure cyber transmission verification highly sensitive data for data protection, privacy and security controls.
5. Assessment of Secure Transmission Data Transmission over Shared Autonomous Networks, Multiple VPNs using same autonomous networks and Shared servicesdedicated lines and secure transmission control.
6. Data Mining of Astute Interception of Shared cyber and VPN traffic for anomaly and malicious patterns for automated responses.
7. Generation of new automatic of cyber security configurations and responses from abstract security policies and control specifications for international data transfer controls.
8. Specifications of security and privacy for international data transfer traceability of autonomous, Regional and US Cyber Networks, Routers, Backbone, VPN, WAN interfaces, including Extranets and Internets.
9. Testing and validation of effectiveness of new approaches against cyber eavesdropping and circumvention security and privacy specifications of highly sensitive data transfer traceability.
10. Development of effective cyber privacy and security controls responses for US cyber transmission and security vulnerabilities and attacks and Interception security and privacy specifications for international data transfer controls.

10.1 Design of Activities

The design of activities to be undertaken include developing new embedded cyber monitoring and filtering detection mechanisms for interesting patterns of data flow, normal and anomaly patterns of data transfers and attack patterns for countermeasures against evasive interception of highly sensitive data, consisting of private, confidential and personally identifiable information for customers, consumers and products at intermediary points of cyber critical infrastructures. This includes specifications for new privacy definitions and control requirements in categories, subcategories and attribute types for secure transmission against interception of highly sensitive data and to ensure that information flow of highly sensitive data in complex systems comply with newly derived security and privacy policies and procedures.

11 Research, Education and Dissemination Timeline

11.1 Year 1: Effective evasive and intelligent attacks facing cyber security and protection

In the first year, research, education and dissemination will consists of analysis, of current effective solutions to the evasive and intelligent attacks facing cyber security and protection. The research consists of developing new and effective embedded monitoring and filtering detection mechanisms, techniques and countermeasures against evasive interception of highly sensitive data, consisting of private, confidential and personally identifiable information for customers, consumers and products at intermediary

points of US cyber critical infrastructures. The research will examine and develop specifications for new privacy definitions and control requirements in categories, subcategories and attribute types for secure transmission against interception of highly sensitive data and to ensure that information flow of highly sensitive data in complex systems comply with newly derived security and privacy policies and procedures. This involves the development of new and effective techniques for validation and verification of effective implementation of these privacy specification controls based on new monitoring techniques for enforcement of security and privacy controls transparency, auditability, traceability accountability and traceability and cyber forensics and breaches assessments and notifications and preventions in critical infrastructures, applications and databases.

11.2 Year 2: Effective design of network and application security techniques and controls cybersecurity

In the second year research, education and dissemination comprises effective design of network and application security techniques and controls for cybersecurity. This involves data acquisition on security parameters and configurations of complex network infrastructure architectures in commercial environments network topologies, interfaces, routing paths, tables, packet filters, access controls, etc. The aggregated logs will include data and application server logs, monitoring tools, including firewalls, IDS, aggregated systems. This will include data mining analysis using normal and anomaly detection using classification, clustering, rule-induction, genetic algorithms to distinguish between anomaly and normal traffic. Detection of patterns of astute attacks and design categories of automated responses to various classes of attack within data-intensive applications over evasive complex infrastructures. This will include analysis of code security vulnerabilities and attacks, prediction of paths of a complex attack and remediation response against such astute cyber attacks using both real-time and off-line analysis and intelligent humans-in-the-loop administrative controls and cross-layer responses for multiple categories, subcategories and features of malicious attacks and subversive attacks. The responses will include efficient and effective combination of automated responses and management of security mechanisms across a complex systems for updating security configurations and parameters and security controls using correlated and synchronized updates in firewalls, intrusion detection, prevention, monitoring and response systems against all classes cyber of attacks. This response will include security mechanisms adapted to the usability and skill levels of multiple user classes: average and skilled users, operators, administrators, forensic experts, research analysts, etc. This will be combined with effective responses to current and emerging complex attacks, using dynamic dispatch of security-enhancement measures and architectures, and integration of response security responses for various thresholds of attacks in multi-protocols of infrastructures hosting data-intensive applications across the Internet.

11.3 Year 3: Managing privacy, confidentiality and identity of highly sensitive data flow on US Broadband and Cyber Networks

In the third year research, education and dissemination will consist of analysis of current effective solutions for managing privacy, confidentiality, and identity in cyberspace for information flow in US and complex systems and the Internet. The research will examine and provide solutions for the dilemma of communication in cyberspace that requires identity and certain degree of disclosure of personal information. This research will analyze broadband networks and web services that have become more pervasive and the types of collection and inference of personal information and the new problems for privacy preservation and the major challenge of security issues and privacy concerns. This includes analysis, research, design and specifications of security, trust and privacy in critical national infrastructures, applications, customer and consumer data, databases, mobile and sensor applications, data min-

ing, web services, digital libraries, in various industries including healthcare, medical, finance, banking, transportation, e-government and e-commerce, etc. The research will examine security and privacy for data-intensive applications, including US and geospatially interconnected systems and broadband networks, where data contents of personal identifiable information, identity class, customer class, etc. impose unique security requirements and introduce trade-offs between security and performance. The responses will include efficient and effective combination of automated responses and management of security mechanisms across a complex systems for updating security configurations and parameters and security controls using correlated and synchronized updates in firewalls, intrusion detection, prevention, monitoring and response systems against all classes cyber of attacks. This response will include security mechanisms adapted to the usability and skill levels of multiple user classes: average and skilled users, operators, administrators, forensic experts, research analysts, etc. This will be combined with effective responses to current and emerging complex attacks, using dynamic dispatch of security-enhancement measures and architectures, and integration of response security responses for various thresholds of attacks in multi-protocols of infrastructures for security, privacy, confidentiality, and identity in cyberspace. The research method will include assessments of real datasets that have been anonymized. The analysis for patterns and classes of private and sensitive data will involve data mining techniques including a combination of various algorithms from rule induction, classification, clustering, genetic algorithms, neural networks and other artificial intelligence techniques etc. These will be used to examine privacy-specific and privacy-related feature attribute sets of categories, classes subclasses and types data requiring unique security and privacy controls and specifications across federated or distributed systems. The datasets will include anonymized datasets from integrated data clusters, applications and databases hosting highly sensitive, confidential data with privacy and security requirements across international borders. The data mining analysis will be applied to feature significant attributes of data classes, subclasses of metadata content fields, sizes, parameters, and type of contents. These will be rated terms in proportion to the risks and requirements for US security and privacy regulatory compliance for both private and public sectors, including multi-national companies. This will result in the design of security control mechanisms includes new security, privacy, on to need-to-validate adequate effective controls for security and privacy. The method for responses to security breaches will include automatic generation of security configurations from policy specifications and validation and verification that information flow in complex systems comply with security and privacy policies. This will include measuring, modeling, analyzing, and validating system trust properties, such as the determination of the extensive efforts required by hackers to defeat security and privacy features.

11.4 Security Specifications for Privacy, Confidentiality and Identity in Cyber Trust for US Cyber standards

The final response and security responses will include specifications for various classes and definitions of Privacy, Confidentiality, and Identity in Cyber Trust for US Cyber standards, and framework development. The final outcome will provide specifications for research areas of naming schemes for privacy and confidentiality; information hiding, anonymity and accountability, transparency; group identity and individual identity, privacy-enabling and disabling technologies; privacy policy monitoring; privacy and privacy-related metrics, integrated adaptable security solutions that combines verification and validation of identity and privacy; formal logics languages, models and methods for specifying and reasoning about privacy. These will include transformation of sensitive data and prevention of its association with particular individuals and utility of the data for research purposes. The transformation methods include generic and application-dependent for deducing ability of transformation methods to preserve utility and prevent the disclosure of sensitive data. Refer to example in Figure 1: Sample Classification Scheme for Transfer of Highly Sensitive Metadata consisting of Personally Identifiable Information Across International

Borders.

11.5 New Classification Scheme of Highly Sensitive Data with Sample

New classification scheme for highly sensitive data for classes, subclasses, and field attributes for metadata and transfers and Master Data Management (MDM) across international borders across critical US infrastructures including Personally Identifiable Information (PII), Non-Public Personal Information (NPI), Health Patient Information (HPI), Customer, Consumer and Product data. See preliminary specifications of Security and Privacy for Highly Sensitive Data in Table 3 and examples of Privacy and Confidentiality attribute classifications in Table 4.

Table 3. New Classification Scheme for International Transfer Highly Sensitive Metadata Confidentiality

New Classification **New Highly Sensitive Metadata Classification**	**Traditional Classification** **Sensitive Data Classification**	**Integrity** **Data Accuracy**	**Privacy and Security** **Privacy and Security Controls**
Strictly Confidential	Strictly Confidential	Very High	Very High
Internal Group Confidential	Confidential	High	Vital
Use Only - Internal	Internal	Medium	Sensitive
Not Applicable	Public Domain	Low	Less Sensitive

Table 4. Examples of New Classification Scheme for International Transfer Highly Sensitive Metadata Confidentiality

Confidentiality	**Categories of MDM Confidentiality Classification**
Permission to View Data	Examples of Customer and Personal Information
Top Confidentiality, Extremely Content Sensitive, Meta Data, Subclass Attributes	Social Security Nos., Drivers Licenses, Credit Cards, Bank Accounts, Passports, Date of Birth
Strictly Confidential	Supervisor IDs, Medical IDs, Unique IDs, etc.
Internal Group Confidential	Home Address, Group IDs,
Use Only - Internal	Company e-mail addresses, Phone
Use Only - Internal	Names, Biographies (Confidentiality based on countries, states)

This will result in specifications, design and development of effective security features for applications, databases and critical infrastructures that transfer highly sensitive data in the presence of attacks more complex and disruptive than those currently observed across US Cyber and interfacing broadband networks to the Internet.

References

[1] M. Asaka, A. Taguchi, and S. Goto. The implementation of ida: Intrusion detection agent system. In *Proceedings of the FIRST Conference for IDA 1999*, Brisbane, Australia, June 1999. http://www.ipa.go.jp/STC/IDA/paper/first.ps.gz.

[2] F. Baboescu and G. Varghese. Scalable packet classification. *IEEE/ACM Transactions on Networking*, 13(1):2–14, February 2005.

[3] T. Bowen, D. Chee, and M. Segal. Building survivable systems: An integrated approach based on intrusion detection and damage containment. In *IEEE Proceedings of the DARPA Information Survivability Conference and Exposition*, volume II of II, pages 84–999. IEEE Computer Society Press, 2000.

[4] Cisco Systems Inc. NetRanger IDS and Software, 2003. San Jose, CA, USA, http://www.cisco.com/univercd/cc/td/doc/product/iaabu/netranger/.

[5] Cisco Systems Inc. Cisco AS5301 Access Server and Software, version 2.28 and release updates, 2005. San Jose, CA, USA.

[6] Cisco Systems Inc. Cisco PIX firewall 525 and Software, version 6.0, 2005. San Jose, CA, USA.

[7] Cisco Systems Inc. Cisco Secure ACS for Windows, version 4.0, 2005. San Jose, CA, USA.

[8] H. Debar and A. Wespi. Aggregation and correlation of intrusion-detection alerts. In *Recent Advances in Intrusion Detection (RAID2001)*, volume 2212 of Lecture Notes in Computer Science, pages 85–103. Springer-Verlag, Berlin, 2001.

[9] K. Djahandari and D. Schackenberg. Coperative Intrusion Traceback and Response Architecture (CITRA). In *Proceedings of the DARPA Information Survivability Conference and Exposition, DISCEX II'01*, volume 1, June 2001.

[10] H. Dreger, A. Feldmann, V. Paxson, and R. Sommer. Operational experiences with high-volume network intrusion detection. In *Proceedings of the 2004 ACM workshop on Formal methods in security engineering*, pages 2–11, Washington DC, USA, 2004. ACM Press.

[11] V. D. Gligor. A note on the denial-of-service problem. In *Proceedings of 1983 IEEE Symposium on Security and Privacy*, pages 139–149, Oakland, CA, USA, April 1983. IEEE Computer Society Press.

[12] S. E. Hansen and T. Atkins. Automated system monitoring and notification with swatch. In *Proceedings of the USENIX Systems Administration (LISA VII) Conference*, pages 145–155, Stanford University, CA, USA, November 1993. USENIX.

[13] J. A. Hartigan. *Clustering Algorithms*. John Wiley and Sons, Inc., New York, USA, 1975.

[14] J. A. Hartigan and M. A. Wong. A k-means clustering algorithm. *Applied Statistics*, 128(3):100–108, July–September 1979.

[15] G. Helmer, J. Wong, V. Honavar, and L. Miller. Intelligent agents for intrusion detection. In *Proceedings of the 2003 IEEE Information Technology Conference*, pages 121–124, Syracuse, NY, USA, September 1998. IEEE Computer Society Press.

[16] R. C. Holte. Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11:63–90, 1993.

[17] R. C. Holte, A. L., and B. W. Porter. Concept learning and the problem of small disjuncts. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 813–818, San Mateo, CA, 1989.

[18] A. hrn. *Discernibility and Rough Sets in Medicine: Tools and Applications*. PhD thesis, Norwegian University of Science and Technology, Department of Computer and Information Science, 1999. http://www.idi.ntnu.no/ aleks/thesis.

[19] K. Iglun. Ustat: A real-time intrusion detection system for unix. In *Proceedings of the 1993 IEEE Symposium on Research in Security and Privacy*, pages 16–28, Oakland, CA, USA, 1993. IEEE Computer Society Press.

[20] C. Iheagwara, A. Blyth, and M. Singhal. A comparative experimental evaluation study of intrusion detection system performance in a gigabit environment. *Journal of Computer Security*, 11(1):1–33, 2003.

[21] Y. F. Jou, F. Gong, C. Sargor, X. Wu, S. F. Wu, H.-C. Chang, and F.-y. Wang. Design and implementation of a scalable intrusion detection system for the protection of network infrastructure. In *IEEE Proceedings of the DARPA Information Survivability Conference and Exposition (DISCEX '00)*, volume 2, pages 69–83. IEEE Computer Society Press, 25–27 January 2000.

[22] G. Koutepas, F. Stamatelopoulos, and B. S. Maglaris. Distributed management architecture for cooperative detection and reaction to DDoS attacks. *IEEE/ACM Transactions on Networking*, 12(1):73–94, March 2004.

[23] C. Kruegel, T. Toth, and E. Kirda. A security policy reinforcement tool for large networks. In *Proceedings of IFIP Conference on Advances in Network and Distributed Systems Security*, Boston, MA, USA, November 2002. Kluwer Academic Publishers.
[24] C. Kruegel, T. Toth, and E. Kirda. Decentralized event correlation for intrusion detection. In *Proceedings of International Conference on information Security and cryptology*, volume Lecture Notes in Computer Science 2288, Berlin, Germany, December 2006. Springer-Verlag, Berlin.
[25] S. Kumar. *Classification and Detection of Intrusions*. PhD thesis, Purdue University, 1995.
[26] S. Kumar and E. H. Spafford. A pattern-matching model for misuse intrusion detection. In *Proceedings of the National Computer Security Conference*, October 1994.
[27] M. V. Mahoney and P. K. Chan. An analysis of the 1999 DARPA Lincoln Laboratory evaluation data for network anomaly detection. In *Recent Advances in Intrusion Detection (RAID2003)*, volume 2820 of Lecture Notes in Computer Science, pages 220–237. Springer-Verlag, Berlin, 2003.
[28] B. Morin, L. Me', H. Debar, and M. Ducasse. M2D2: A formal data model for IDS alert correlation. In *Recent Advances in Intrusion Detection (RAID2002)*, volume 2515 of Lecture Notes in Computer Science, pages 115–137, Zurich, Switzerland, 16–18, October 2002. Springer-Verlag, Berlin.
[29] Network Associates. Blackice IDS, 2003. Herndon, Virginia, USA, http://www.blackice.iss.net.
[30] Network Associates. Internet Security Systems (ISS) Real Secure IDS, version 6.5, 7.0, 2004. Herndon, Virginia, USA, http://www.iss.net.
[31] Network Associates. McAfee Intrushield IDS: 4000 Series, 2005. Santa Clara, CA, USA.
[32] Network Associates. McAfee Intrushield IDS: 4000 Series, 2007. Santa Clara, CA, USA.
[33] Network Flight Recorder. Network Flight Recorder (NFR) IDS, 2001. http://www.nfr.net/.
[34] V. Paxson. Bro: A system for detecting network intruders in real-time. *Computer Networks*, 31(23–24):2435–2463, 1999.
[35] K. L. Petersen. IDA: Intrusion Detection Alert. In *Proceedings of the IEEE Annual International Computer Software and Applications Conference*, pages 306–311. IEEE Computer Society Press, September 1992.
[36] L. Portnoy, E. Eskin, and S. Solfo. Intrusion detection with unlabelled data using clustering. In *Proceedings of ACM CSS Workshop on Data Mining Applied to Security (DMSA-2001)*, pages 76–105, 2001.
[37] T. H. Ptacek and T. N. Newsham. Insertion, evasion and denial of service: Eluding network intrusion detection. Technical Report, Secure Networks (McAfee) Inc., Santa Clara, CA, USA, January 1998. http://citeseer.ist.psu.edu/ptacek98insertion.html.
[38] M. J. Ranum, K. Landfield, M. Stolarchuk, M. Sienkiewicz, A. Lambeth, and E. Wall. Implementing a generalized tool for network monitoring. In *Proceedings of the Eleventh Systems Administration Conference (LISA'97)*, San Diego, CA, USA, October 1997.
[39] M. Roesch. Snort: Lightweight intrusion detection for networks. In *USENIX Lisa-99*, pages 229–238. USENIX, 1999.
[40] Rulequest Research. Rule Induction with C5.0, See5/Cubist software, 2005.
[41] S. R. Snapp, J. Brentano, G. V. Dias, T. L. Goan, L. T. Heberlein, C.-l. Ho, K. N. Levitt, B. Mukherjee, S. E. Smaha, T. Grance, D. M. Teal, and D. Mansur. DIDS (Distrubted intrusion detection system): Motivation, architecture and an early prototype. In *Proceedings of the 14th National Security Conference*, pages 167–176, October 1991.
[42] SPSS Inc. SPSS Statistical Software, 2007. Chicago, IL, USA, http://www.spss.com.
[43] S. Staniford-Chen, S. Cheung, R. Crawford, M. Dilger, J. Frank, J. Hoagland, K. Levitt, C. Wee, R. Yip, and D. Zerkle. Grids: A graph-based intrusion detection sysyem for large networks. In *Proceedings of the 20th National Information Systems Security Conference*, volume 1, pages 361–370. National Computer Security Centre Press, October 1996.
[44] Symantec Corporation. Norton Internet Security and Symantec Host IDS, 2006. Cupertino, CA, USA, http://www.symantec.com.
[45] T. F. Toth. *Improving Intrusion Detection Systems*. PhD thesis, Technical University of Vienna, 2003.
[46] Touch Technologies Inc. Touch Technologies INSA IDS, 2006. San Diego, CA, USA, http://www.ttinet.com/.
[47] G. Vigna, S. Eckmann, and R. A. Kemmerer. The stat tool suite. In *Proceedings of DISCEX 2000*, pages 1046–1053. IEEE Computer Society Press, January 2000.
[48] G. Vigna and R. A. Kemmerer. Netstat: A network-based intrusion detection system. In *Proceedings of the 14th Annual Computer Security Applications Conference*, December 1998.
[49] S. Vinterbo and A. hrn. Minimal approximate hitting sets and rule templates. *International Journal of Approximate Reasoning*, 25(2):123–143, 2000.