Proceedings of the 12th Colloquium for Information Systems Security Education
University of Texas, Dallas
Dallas, TX June 2 - 4, 2008

Security Visualization Survey

Denise Ferebee and Dipankar Dasgupta, *University of Memphis*

Abstract – Visualization plays a major role in understanding and interpreting security requirements. Security visualization means different things to different people. Some consider it as viewing the state of the environment and system. The purpose of this paper is to review some of the current methods used in security visualization.

Index terms – information security, human intuition, administrator experience, visualization, visual analysis

I. Introduction

Computer networks are becoming larger and even more complex. This stems from the increase in technological advances such as mobile/wireless devices, business expansions, and the world becoming more connected requiring more infrastructure. Therefore, when issues occur (i.e. denial-of-service, authentication failures, data integrity issues, viruses, etc) quick analysis of large amounts of data is required. Security analysis may require reviewing logs, reviewing the current state of network and/or server equipment, collaboration between multiple workgroups to form a model of the security event, etc. Hence, the time needed for the analysis needs to be minimized in order to provide a quick mitigating response.

Cognition is defined as "the act or process of knowing; perception". It is accomplished by interacting with cognitive tools, pencil and paper, computer-based intellectual supports, etc. As an old saying goes "a picture is worth a thousand words". In other words, you are able to convey more meaning from a picture than a written explanation. Hence, the purpose of information visualization is to convey meaningful information in reference to potentially large amounts of data [1]. Therefore, visualizations have become an important part in security analysis systems.

Visualizations provide advantages such as but not limited to the following [1]:

- Allows emergent properties that were not anticipated to be perceived
- Problems with the data becomes apparent
- Understanding of large-scale and small-scale features is facilitated
- Hypothesis formations are facilitated

The question becomes "what purpose does this serve when it comes to security". Security event data is massive and spans several devices requiring collaboration from each device administrator. Visualization may provide a means to improve the correlation process. Thus, this paper covers the following stages of the visualization process [1]:

- Data collection and sanitization
- Data transformation and event correlation
- Creation of the visualization of results

Each of these steps will take you further to creating a meaningful visualization.

II. Problem Definition

Over the years, computer usage has increased due to the attainability of equipment, the decrease in internet access costs, and the makeup of the user community. However, it was not without issue. As usage increased, so did crime. Computer crime is classified into one of the following categories [2]:

- A computer is used to commit a crime
- A computer is the target of a crime

Fred Cohen in 1983 created a computer virus that acquired privileges on a VAX-11/750 running the UNIX operating system. This was one of many viruses that ended the complacent or trusting nature of people toward general technological openness [3]. Other issues for users can be one of but not limited to the following [4]:

- Identity theft
- The randomness of the computer attacks
- Economic impact of clean-ups after misuse
- Difficulty of mitigating misuse

Because of rapid technological changes, the tools used for mitigation have to be adaptable. Many of the security tools fall under the following category:

Center for Information Assurance
Department of Computer Science
University of Memphis, Memphis
Denise Ferebee email: dferebee@memphis.edu
Dipankar Dasgupta email: dasgupta@memphis.edu

- Firewalls – personal and network
- Intrusion Detection and Prevention Systems
- Virus Scanning Tools
- Integrity Checkers

These tools generate a large amount of data which needs to be analyzed for security purposes. In the next section, we will cover the tasks and issues in reference to the collection and sanitization process as the first step in information visualization.

III. Data Collection

As stated before, there is an abundance of security data requiring review in order to mitigate responses. This typically requires collaboration between various groups of people (e.g. network administrators, system administrators, application owners, Internet security groups, etc.) Because, each of these groups collect their own event information, share data, and correlate events; standard (i.e. agreed upon) data collection and validation methods must be adhered. This prevents security breaches by not disclosing too much information (i.e. organization specific information to outside sources). Therefore, this section will cover some of the current methods for data collection and sanitization for event collaborations. This data will eventually be correlated in order to provide a meaningful visualization.

A. Third-Party Processing

What does data collection have to do with information visualization? The answer to this question is a basic one. In order to have visualization, it must be based on some data that can be mined in a visual manner in order to provide comparisons and correlations that are normally not seen.

Many organizations or companies either contribute data to in-house collaboration efforts where they are used with tools like Arcsight's Interactive Discovery [5] and Snort's Basic Analysis and Security Engine (BASE) [6] or they contribute security logs to external cooperatives such as DShield and DeepSight [7, 8]. These contributions facilitate real time warning services of security attacks. However, data collection in this manner and in general is not without challenges.

Data provided to external cooperatives can be a potential information leakage point. Network topology is revealed via fingerprinting along with the affects of the current attacks revealing vulnerabilities. Also, false information can be given which in turn skews the analysis of the attack pattern. Hence, in order to provide data and participate in these cooperatives; many rules must be adhered such as but not limited to the following [7]:

- False information must not be contributed
- Data must not be used for malicious intent

As previously mentioned, these guidelines prevent the skewing of the credibility of the information used in security forecasts. Therefore, how contributors are authenticated and the requirements for data submission are important.

1. Authentication of Data Contributors

In order to guarantee that the individuals that are providing the data are who they state they are, various authentication mechanisms can be employed. These mechanisms range from username and password and anonymous credentials to cryptography. The major issue with authenticating the contributor is that they divulge their company or organization information [7]. In today's society this can be detrimental to a company or organization in the aspect of bad press and monetary loss. This leads to a problem which will be discussed in the next section called privacy preserving data mining or the anonymity problem. [14]

Next, let us consider encryption (i.e. public key / private key). This is a very good way of mitigating faked submissions. However, it inadvertently creates the problem of key management. Keys have to be stored in a secure location and who will manage it becomes an issue [7, 2]. However, these authentication methods do not guard against the attacker being a legitimate contributor [7]. So, the question becomes, does authenticating the contributor buy anything? Yes, it puts at least one layer of security in place to prevent misuse.

2. Submission Requirements

Many collaboration groups have specific data requirements. DeepSight requires an installation of their collection tool called DeepSight Extractor. This tool can collect data from some of the following firewalls or IDSs: NetScreen, Checkpoint Firewall, ZoneAlarm, etc [7,8].

Other collaboration efforts may require data to be in a specific format. This provides an agreed upon nomenclature for communicating event properties and courses of action. This will be explored more in the data sanitization section. Next, we will look at using network sensors for data collection as it is applied to network intrusion detection.

3. Sensors

Because of the sheer volume of data, research was done by Radford University and Eastman Chemical Company using network sensors such as a Gumstix [9] for data collection. The purpose was to use network sensors to collect intrusion data in order to offload the resource requirement from the main network and send it to management application for analysis.

Their concept called a Security Monitoring System for Information Assurance, Analysis, and Survivability of Network Hazards (SMASH) is based on combining the jobs of a network sniffer and automatic response. Basically, if one network sensor detects an attack and all sensors block the attack before it occurs at other sensor locations. Basically, the network sensor is another device that can be leveraged for providing network intrusion detection data whether it is detection or response.

4. In-House Applications

Some organizations collect data to use with open source or purchased commercial tools in-house. Data is used in tools like Arcsight's Interactive Discovery provides the capability of out-of-the box visual perspectives. It includes visual charts such as parabox, time slice, histogram, and scatter plot [5]. It allows for pan and zoom and other rich visualization features. Figure 1 is a sample screen showing traffic profile by protocol, the number of sensors, alerts, etc. [6]. Finally, Figure 2 is a sample screen showing a customizable view of security incident data such as alert states, severity, etc. [10].

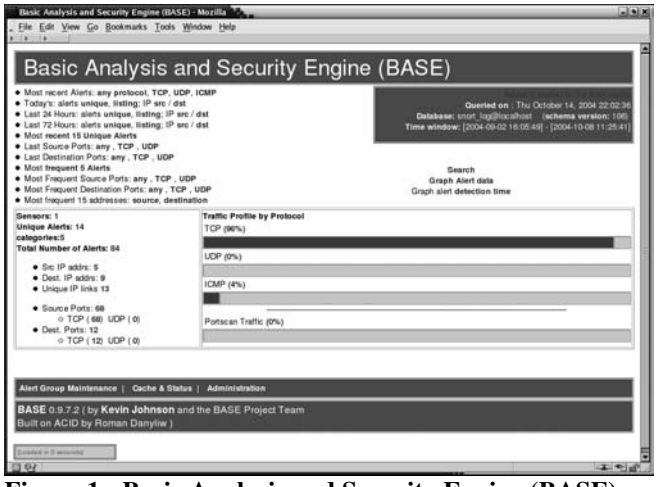

Figure 1 - Basic Analysis and Security Engine (BASE) [6]

No matter whether you use an in-house or commercial tool or an external collaboration group, understanding what is needed for data collection is important.

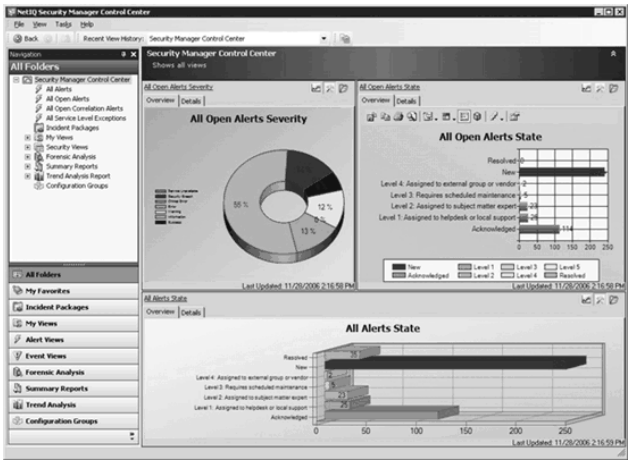

Figure 2 - NetIQ Security Manager [10]

B. Sanitization

In this section, data sanitization will be discussed. The obvious question is what does this have to do with data visualization. As discussed in Section 2.2, companies or organizations send their security data to cooperatives to be used in predicting security events across the Internet. Some of these cooperatives provide visualizations in reference to the health of the internet. Also, some of the data is used by contributors for comparative analysis purposes.

1. Data Formats

When data is not in a similar format and similar classifications, it will be difficult to classify or correlate events. Different tools and collaboration organizations require specific data formats. These data formats can consist of Intrusion Detection Message Exchange Format (IDMEF) [11] to appliance/application specific such as Checkpoint Firewalls or Cisco PIX [7,8].

We will look at Incident Object Description and Exchange Format (IODEF). It is an Extensible Markup Language (XML) format for facilitating collaboration. Also, it was created with the forethought of being used for building incident correlation and statistical systems. IODEF is based on IDMEF providing interoperability [12, 20].

Incident Class format is able to provide a free-form description, characterizations of an incident, detection time, start time, attack method, etc. This class has four attributes purpose (i.e. why the document was created), ext-purpose, language, and restriction (i.e. the requested security measures for disclosure of the information contained in the document) . There are other classes that

are part of this specification that go beyond the scope of this document [13].

2. Security Issues

Now that data has been collected, how do we sanitize it to prevent information leakage yet provide enough information for a clear understanding of the events that are occurring? In order to tackle this problem, let's first look at the issues of not sanitizing the data. They consist of but are not limited to the following [7]:

- **Privacy issues** – disclosing network topologies and security mechanisms
- **Probe response attacks** – triggering rare rules in signature-based intrusion detection systems (IDSs), employing rare ports combinations, etc. that can be later recovered
- **Fingerprinting** – searching through patterns in the data to identify hosts or devices
- **Data accuracy** – due to the vast amount of data and collaborative analysis precision and accuracy is scrutinized
- **Usability issues** – security conflicts with usability in the aspect of obfuscating the IP addresses that appear in alerts
- **Blending attacks** – submission of fake records in the hopes that it produces a recognized pattern in released data sets

By exposing information in reference to network topology and security mechanisms, it opens volatile information about what types of equipment is in place and provides metadata that can be used to circumvent security mechanisms [7]. Next, by triggering rare rules in IDSs, it gives the attacker information in reference to the effectiveness of their attacks. Finally, submission of false records skews the accuracy of the analysis performed on the data making users mistrust the effectiveness [7]. Therefore, in the next sections we will discuss how to mitigate against these issues/risks.

3. Privacy-Preserving Data Mining

One of the fore mentioned issues with event cooperatives or collaborations is that the data provided is a potential information leakage point. This may potentially reveal information about the organizations network that could be detrimental. Therefore, this section focuses on privacy-preserving data mining techniques [7, 14]. The two areas that will be focused on are homomorphic cryptography [14] and IP virtualization modeling [7].

Before discussing homomorphric encryption, let us define homomorphism. In set theory, homomorphism is the replacing of a symbol S with a string W (i.e. h(s) = w).

The approach taken for preserving privacy was to use homomorphic encryption and digital envelope techniques. This involves using privacy-oriented protocols with a homomorphic additive to maintain privacy while mining data [14].

The other privacy preserving technique involves research in IP virtualization. This involves maintaining the network topological information while not revealing the actual IP address contained in the data. A challenge to this is if an attacker is able to get determine some portion of the mapping between real and virtual addresses. They may be able to determine all of the IP addresses invalidating the privacy feature. Further, research in this area is needed [7].

C. Automated Data Processing

Automated data processing employs techniques that provide immediate response to security events that typically take place before an administrator's intervention. An example of this is a solution such as a proposed solution that makes use of artificial immune type of response [4, 15, 16].

Let us take a look at the artificial immune solution. This system should be multilayered by which several mechanisms (e.g. virus scanners, firewalls, etc) communicate by message passing during the immune response phase in order to guard the network against and respond to foreign pathogens (i.e. in order to prevent security breaches) [15, 16]. It should base its decision on multiple signals to help differentiate between dangerous or harmless intruder minimizing false positives. The major advantage of this technique is that the immune system learns what is self and not self, works on a distributed level preventing single point of failure, and can launch threat and attack responses [4, 15, 16]. This type of artificial immune system (which is mainly based on a negative selection algorithm) was applied to anomaly visualization by using self-organizing maps (SOM) to produce a 2-dimensional map that represents the feature spaces [15]. Because of the self/non-self recognition of the nodes are able to be classified as normal (black), unknown anomaly (gray), and known anomaly (white) and labeled accordingly [15].

In the next few sections, we will pick backup with event correlation and visualization that do not make use of these techniques.

IV. EVENT CORRELATION

So far, data collection and sanitization issues have been covered. In this section we will discuss how event correlation gets us closer to information visualization.

The main issue with mitigating security risk is to make sense out of the information that has been provided. The average security engineer, system administrator, and developer have to wade through mountains of data in order to determine what events are occurring. Therefore, this section will cover some of the current techniques used to transform, categorize, and make sense of the data.

First, let's discuss some of the disparate systems or appliances from which data is provided. Data can be provided from firewalls, virus scanners, applications, computers, etc. Each of these systems has their own set of rules to determine and alert on misuse. This can be either by sending an email to alert on a message in a log file, a pop-up on the screen, etc. However, they may not have communicated this alert to one of the fore mentioned systems or appliances. Therefore, requiring the administrator, network engineer, or developer to review each system separately and make a guess as to what may be happening. In the next section, we will cover data transformation and event modeling and how they help us to make some sense of the data.

A. Data Transformation

In many aspects of making sense out of data, there is a tendency to map it into a classification. Figure 3 example of a type of transformation is the Canonical Event. It is a cross-correlation of heterogeneous events in the aspect of event correlation for forensics (EFC). This type of correlation focuses on post hoc (i.e. after an event has occurred) [17].

Each event has the following required attributes [17]:

- EventID – unique identifier
- Time – time of the event
- EventType – type of event
- Result – success, failure, etc.

Through an extensive scenario matching algorithm events are matched into Logical Event Patterns (LEP). This matching is controlled by a XML configuration file which contains <patterns> element containing all of the LEPs. Then, a scenario matching algorithm is applied by building a custom SQL query [17].

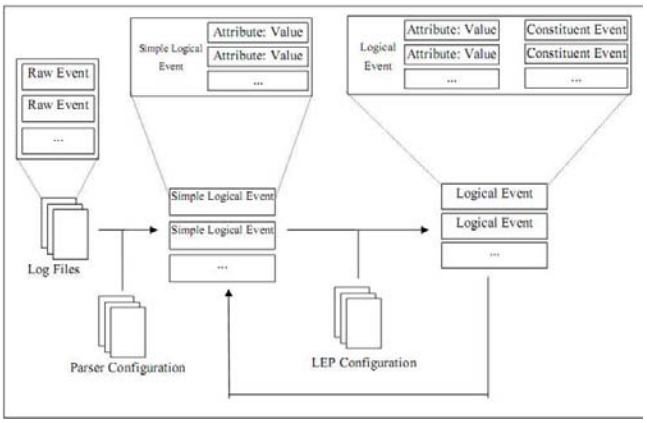

Figure 3 - Auto - ECF system [17]

By mapping the data to a classification, one can create a visualization based on that mapping that is meaningful.

B. Event Modeling

In this section, we will discuss event modeling. The question of "What does event modeling have to do with security visualization?" comes to mind. If you think about it, what you are showing is a relationship in the data that is being visualized. In order to visualize data you have to understand the data and the meaning it is trying to convey.

Some of the types of event modeling used deal with the following techniques:

- Snort Classifications [18]
- Anomaly detection [4, 15,16]
- Negative selection [4, 15,16]
- Function based [19]
- Link analysis [20]
- Machine learning [4]

In this section, we will cover a few of the fore mentioned techniques. In anomaly detection, the model of self and non-self is used. This provides a way for determining the difference between a normal and abnormal state (i.e. a state for which is not a normal system behavior) [4, 15, 16].

Link analysis deals with extracting useful information from large data sets of associations. Its visual representation consists of a showing a graph of entity associations. The associations are achieved via machine learning to detect patterns. Also, Signatures from Snort can be used to classify types of events that occur [18]. Creating a model of the security events that are occurring depends mainly on the relationships that need be shown in reference to the data being analyzed.

V. VISUALIZATION IN SECURITY TOOLS

So far, we have looked at data collection, sanitization, and event correlation in the process of creating visualization. This section will be dedicated to looking at some of the visualization characteristics along with examples of open source, academic, and commercial packages.

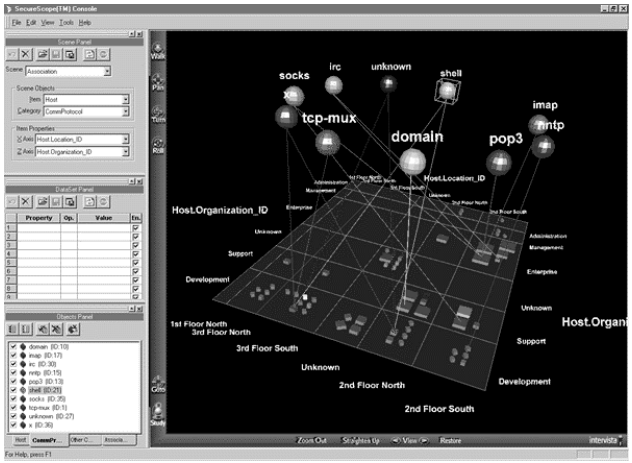

Figure 4 – SecureScope Sample Scene [21]

A. Characteristics

When it comes to visualization, many factors in reference to the data and the user play a part. These can range from having a user that is color blind to trying to show blood flow on an image of the heart. In visualization entities (i.e. heart, blood, etc.) would be the objects that we would like to visualize along with the structures or patterns that relate them. Properties of entities and relationships are called attributes when they cannot be separated from an entity (e.g. separating temperature from water) [1].

Attributes have four levels of measurement. They consist of nominal (category), ordinal (integer), interval (real-number), and ratio (real-number) according to S.S. Stevens [1]. These are very important when it comes to visualization technique discussions. An example would be using a pie chart to represent nominal/category data because we tend to interpret quantity as size. There are other things to consider when creating visualization such as: luminance, brightness, contrast, color, etc. that go beyond the scope of this document.

B. Example Applications

Some of the visualizations featured in this section are from NetIQ [10] as shown in Figure 3 and Figure 5 and M and R Security Console [22] as shown in Figure 6. First, we will look at NetIQ and some of its features. NetIQ provides a distributed log warehouse, rule based correlation, and agent based change detection at the host/system level. In Figure 5, a quantitative representation is being shown of the number of events by using a bar chart.

Next, SecureScope is a three-dimensional visualization tool specializing in network security event data [21]. Figure 4 shows the relationship between service and organization and location. In Figure 6, the M and R Security Console is also using a quantitative representation of the Episode Frequencies that are occurring [22]. Figure 7 is a sample screen of the Attack Path and Layer 2 Mitigation [23].

Each of these examples shows some relationship between in the data that they represent. These relationships help people to explore and discover new information. Therefore, understanding how data relates and how to take advantage of visual perception with color, symbols, luminance, etc. help to make affective visualizations.

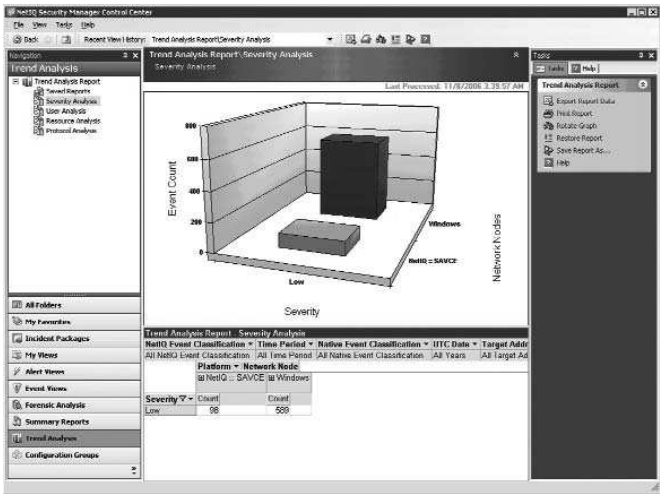

Figure 5 – NetIQ Security Manager [10]

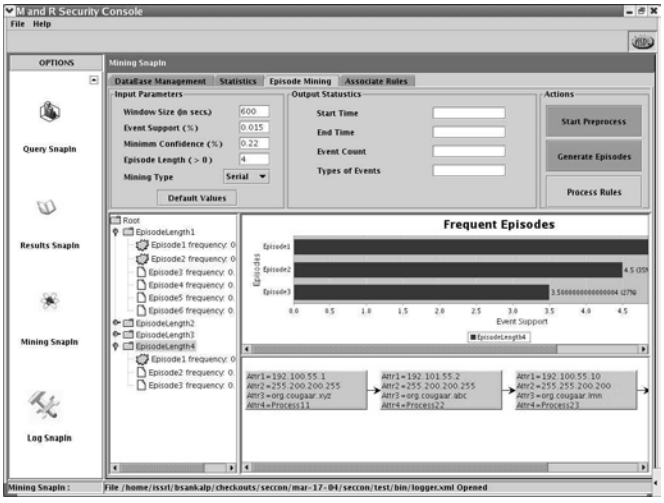

Figure 6 - M and R Security Console - Alert Miner [17]

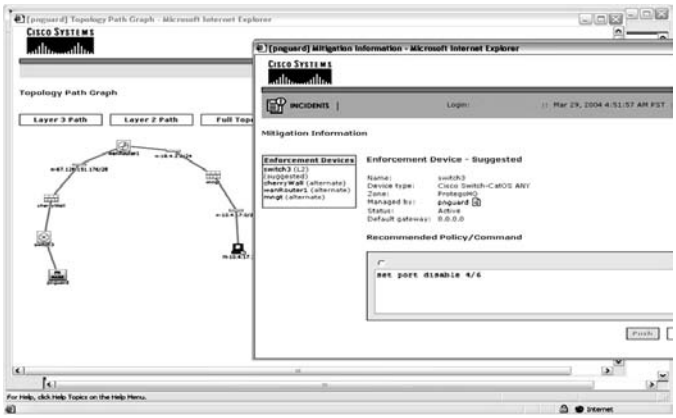

Figure 7 – CISCO'S Mitigation, Analysis, and Response System (MARS) [23]

VI. SUMMARY

This paper has covered a number of issues related to security event visualization particular data collection sanitization, processing and displaying meaningful information in a precise way. Without proper visualization it is hard to find meaning in the data because of the sheer volume.

The purpose for this survey was to determine what is currently being done to meet the information security visualization need in order to improve the incident response process. In this survey security concerns in reference to user authentication, privacy, and correlation were covered. It was observed that without preserving privacy that it may cause loss of confidence and money when topology specific data is exposed while trying to resolve a security event via external collaboration. However, there is risk with almost everything. This paper covers examples that illustrate security visualizations that show relationships between security mechanisms and events. If we pay attention to the mitigation maybe additional visualizations can be created that will lessen the security response time and cost.

VII. REFERENCES

[1] Ware, C. *Information Visualization: Perception for Design*. San Francisco, CA: Morgan Kaufmann Publishers Inc., 2004.

[2] Kruse II, W. and Heiser J. *Computer Forensics Incident Response Essentials*, Indianpolis, IN: Addison Wesley, 2002.

[3] Bishop, M. *Introduction to Computer Security.* Boston, MA: Addison-Wesley, 2004.

[4] Vemuri, V. (ed.), Enhancing Computer Security with Smart Technology, CRC Press, 2005.

[5] Arcsight Home Page < http://www.arcsight.com>, March 2008.

[6] Basic Analysis and Security Engine (BASE) Home Page <http://base.secureideas.net/>, March 2008.

[7] Porras, P. and Shmatikov, V. 2007. Large-scale collection and sanitization of network security data: risks and challenges. *In Proceedings of the 2006 Workshop on New Security Paradigms* (Germany, September 19 - 22, 2006). NSPW '06. ACM Press, New York, NY, 57-64.

[8] Symantec – DeepSight Analyzer Home Page <http://aris.securityfocus.com/>, March 2008.

[9] Derrick, E. J., Tibbs, R. W., and Reynolds, L. L. 2007. Investigating new approaches to data collection, management and analysis for network intrusion detection. *In Proceedings of the 45th Annual Southeast Regional Conference* (Winston-Salem, North Carolina, March 23 - 24, 2007). ACM-SE 45. ACM, New York, NY, 283-287

[10] NetIQ Home Page <http://www.netiq.com/products/sm/default.asp>, March 2008.

[11] Cover Pages: IDMEF Home Page <http://xml.coverpages.org/idmef.html>, April 2008.

[12] Cover Pages: IODEF Home Page <http://xml.coverpages.org/iodef.html>, April 2008.

[13] Internet Engineering Task Force<http://www.ietf.org/>, April 2008

[14] Zhan, J. and Matwin, S. 2006. A Crypto-Based Approach to Privacy-Preserving Collaborative Data Mining. In *Proceedings of the Sixth IEEE international Conference on Data Mining - Workshops* (December 18 - 22, 2006). ICDMW. IEEE Computer Society, Washington, DC, 546-550.

[15] González, F. A., Galeano, J. C., Rojas, D. A., and Veloza-Suan, A. 2005. Discriminating and visualizing anomalies using negative selection and self-organizing maps. In *Proceedings of the 2005 Conference on Genetic and Evolutionary Computation* (Washington DC, USA, June 25 - 29, 2005). H. Beyer, Ed. GECCO '05. ACM, New York, NY, 297-304.

[16] Dasgupta, D. Immuno-Inspired Autonomic System for Cyber Defense. Information Security Technical Report,Elsevier Ltd,. 12 4 December, 2007

[17] Abbott, J., Bell, J., Clark, A., De Vel, O., and Mohay, G. 2006. Automated recognition of event scenarios for digital forensics. *In Proceedings of the 2006 ACM Symposium on Applied Computing* (Dijon, France, April 23 - 27, 2006). SAC '06. ACM Press, New York, NY, 293-300.

[18] Hideshima, Y. and Koike, H. 2006. STARMINE: a visualization system for cyber attacks. *In Proceedings of the Asia Pacific Symposium on information Visualisation - Volume 60* (Tokyo, Japan, February 01 - 01, 2006). K. Misue, K. Sugiyama, and J. Tanaka, Eds. ACM International Conference Proceeding Series, vol. 243. Australian Computer Society, Darlinghurst, Australia, 131-138.

[19] Treinish, L. A Function-Based Data Model for Visualization. <http://www.research.ibm.com/people/l/lloydt/dm/function/dm_fn.htm>, 2008

[20] Yin, X., Yurcik, W., Treaster, M., Li, Y., and Lakkaraju, K. 2004. VisFlowConnect: netflow visualizations of link relationships for security situational awareness. *In Proceedings of the 2004 ACM Workshop on Visualization and Data Mining For Computer Security* (Washington DC, USA, October 29 - 29, 2004). VizSEC/DMSEC '04. ACM Press, New York, NY, 26-34.

[21] Secure Decisions <http://www.securedecisions.com>, April 2008.

[22] Dasgupta, D., Rodriguez, J., Balachandran, S. Efficient Visualization of Security Events in a Large Agent Society. *Conference on Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security*. Orlando, Florida: March, 2005

[23] CISCO Home Page <http://www.cisco.com>, April 2008.