

# Can Management Predict Information Security Threats by Utilizing a Data Warehouse?

Philip Kim

[Pxkst1@mail.rmu.edu](mailto:Pxkst1@mail.rmu.edu)

Lee Jonathan Steen

[Ljsst8@mail.rmu.edu](mailto:Ljsst8@mail.rmu.edu)

Computer Information Systems, Robert Morris University  
Coraopolis, PA 15108, USA

## Abstract

In many organizations, management has the responsibility of implementing information security countermeasures to detect, minimize, and defend against information security threats. Most of these countermeasures traditionally adopt a passive approach to securing corporate data. This paper proposes a new theoretical framework for management to utilize an information security data warehouse to identify security breach patterns, in order to predict when potential breaches are most likely to occur, thus taking a more proactive approach to securing information assets.

**Keywords:** Information Security, Security Management, Predictive Security Model, Security Breaches, and Secure Data Warehousing.

## 1. INTRODUCTION

Information security has experienced exponential growth and consideration in recent years. Information has become a major financial staple for organizations. Boisot (1998) explains that organizational information and corporate data are the new currency of business in the Information Age. And in light of the current worldwide economic state, according to the CSI Computer Crime & Security Survey (2008), the cost of information security breaches is increasing. Surprisingly, security spending is one area that organizations are beginning to cut back on within their information technology budgets. This gives managers all the more reason to be even more vigilant in protecting systems as security threats and breaches will continue to increase.

The typical countermeasures for hardening, or securing information systems include layering defense systems, diversifying defense methods (Cole, Krutz, & Conley, 2005) ongoing management of hardware and soft-

ware updates or patches, access controls with auditing (Gallegos et al., 2004), and authentication mechanisms (Soper, Demirkan, & Goul, 2007). The issue with traditional information security countermeasures is that organizations often adopt a "passive" approach. This passive approach is operationally defined as implementing security policies and countermeasures, and hope they work. Management can gain a false sense of assurance because they purchase the newest security systems, and yet even with the most advanced security technologies in place, breaches still occur.

All organizations that connect to the Internet recognize the need to install some type of firewall. And yet simply deploying a firewall or several firewalls and not securing any further presents a single point of failure (Ranum, 1993). Other companies put up a myriad of defense mechanisms in place to prevent attacks without knowing what types of attacks will occur or even know when previous attacks *have occurred*. Other firms try to stop or mitigate the attack happening in

real time (e.g. Intrusion Detection Systems (IDS), or review logs and audits to see when attacks or breaches occurred). Even the IDSs can be viewed as a passive approach to security because they often require an IDS Analyst to detect and respond to an attack, which is a reactive response.

This paper presents a modified approach from the traditional passive and reactive security countermeasures to a more proactive and risk-based approach to manage information security risks. The proactive and risk-based defense has two parts. First is predicting when and where an attack is likely to occur. Second is to implement a risk-based methodology of safeguarding the most critical assets against an attack. This proactive approach allows an organization to actively look for where breaches are likely to occur. Once this is known, a risk-based approach allows the organization to focus more resources on the high value and more vulnerable areas. This makes the use of security resources more efficient and results in a better return on investment. The focus of this paper will be to discuss how management can potentially predict where security attacks will most likely occur and the significance of utilizing a risk-based approach to determining a corporation's critical assets.

The predictive security model presented in this paper for predicting security breaches involves the use of a data warehouse. The research will review current security threats, data warehousing, and present a predictive model. The reader is encouraged to critique, adapt, and modify the model to their specific organization or industry. The paper concludes with the potential gaps within the model and topics for future research.

## 2. SECURITY THREATS

A security threat is any circumstance or event with the potential to cause harm to an asset (White & Conklin, 2008). The most common threat for organizations is not only from outside attackers trying to gain unauthorized access, but inside employees who already have access and intimate knowledge of the systems. Internal employees pose a significant threat not only due to internal access rights, but often because of simple human error. Unwitting internal employees who are untrained in the proper use of sys-

tems and corporate security procedures are a threat to the information security environment (Cole, Krutz, & Conley, 2005).

Attacks can be triggered from the outside when the unauthorized party (attacker) is seeking access into the corporate network. A successful breach could result in loss of information or worse, the spread of sensitive information. Attackers can use stolen information to sell to the highest bidder, or use the data to negatively affect the organization's reputation, or possibly some other personal gain. Attacks can be initiated from the inside for the many of the same reasons including, employees seeking to profit from confidential or proprietary data, and disgruntled employee(s) seeking revenge against a co-worker or employer. Another growing threat is social engineering attacks, which can be successful when employees are not trained properly and have little to no understanding of corporate security policies (Mittnick & Simon, 2002).

The first line of defense towards mitigating security threats are often enforced by standard policies and procedures. The policies and procedures should address how company technology and information assets are to be controlled and handled. Corporate policies and procedures should also include training employees on system and network updates for software and hardware, physical security practices, and updated regulatory guidances (Samuelle, 2008).

## 3. DATA WAREHOUSING AND DATA MINING

A data warehouse is a repository of large amounts of data in a single, non-normalized location. The data warehouse gets its information from databases, transactional systems, data marts, and from other sources in the organization where data is generated (Agosta, 2000; Simon, 1997).

Data warehouse information can be analyzed for patterns that emerge and for useful information by using data mining techniques. These techniques can use mathematical analyses and algorithms to discover patterns that are not easily identified by manual review (Zarsky, 2003).

In order for a data warehouse to be most effective, it has to contain relevant and up-to-date data that can lead to useful information. For the purpose of this theoretical framework, we are most interested in how a dedicated information security data warehouse can be a central repository for an organization's history or previous experience with security threats, attacks, and breaches. That is, the information security data warehouse will be a storehouse of any security breaches that have occurred and all of the possibly related information to those security breaches. But gathering all data available may be too time-consuming and may prove to be unhelpful if it is not relevant.

It is worth analyzing applications within security that have built themselves around the use of data warehousing and/or data mining to fully understand the implications of the new idea being presented. These applications reviewed include intelligent intrusion detection systems (IDS) (Helmer et al., 1998; Lee et al., 2001), internet protocol traffic and measurement analysis (Caceres et al., 2000), identification of fraud within telecommunications (Cao et al., 2004), breach propagation detection for knowledge-sharing within an organization (Soper, Demirkan, & Goul, 2007), and role identification for security administration (Kuhlmann, & Schimpf, 2003).

#### **4. ADAPTATION OF CURRENT APPLICATIONS**

Intelligent IDSs are designed to identify patterns within a computer network that appear to be inconsistent with a baseline and analyze the pattern to determine if an intrusion is occurring or has occurred. Real-time IDS systems need to be calibrated to reduce false-positives, or false alarms. IDS systems must also be efficient so they do not bog down network resources (Lee et al., 2001). Although IDS systems are considered effective in detecting security anomalies, they do not have the capability to anticipate breaches.

Lee, et al. (2001) states that "[t]o improve accuracy, data mining programs are used to analyze audit data and extract features that can distinguish normal activities from intrusions" (p.89). By utilizing data mining techniques, it is possible to detect breaches in

real time. It would be possible to take the current data mining process, store the same audit information in a data warehouse and perform queries to determine the possible emergence of intrusion patterns. Organizations may be able to observe intrusion patterns by harnessing the ability to store, process, and query large amounts of information within the data warehouse. As more data are collected and intrusion patterns are identified, management could perceivably determine causal or corollary relationships to emergent security intrusions.

Another type of IDS uses data gathering agents to collect system logs and activity. The IDS then gathers and summarizes the data into an easy-to-understand common format (Helmer et al., 1998). Using this data, lower level agents classify the data and send it up to a higher level. What this means is at a low level, several agents analyze the data and summarize it by a common mean. The common mean summary (the summarized data) is then passed one level higher. The higher level receives several summaries together and analyzes them, common means them again (summary of previous summarized data put together), and passes the data to the next higher level. This process is done until a single level makes a decision for the whole network. Adapting this to the information security predictive model would mean starting with the lowest level of security breach data, summarizing the data into meaningful information, and passing it up higher until it reaches the level of decision. The security breach data would need to be accessible and summarized at each level to determine if patterns exist within the breach (or intrusion).

Another area that is similar in design, measurement and analysis of data is the monitoring of internet protocol (IP) network usage and behavior. A current system of monitoring and measuring IP network usage and behavior is being used by AT&T (Caceres et al., 2000). This study describes how IP traffic is stored in a data warehouse and analyzed by data mining techniques to understand the behavior of IP traffic in a network. This type of data could be instrumental to understanding when a breach is likely to occur. If a specific segment of a network is experiencing unusually high traf-

fic and the traffic begins to migrate, a pattern of movement may be emerging. Utilizing the data warehouse's real-time monitoring of IP traffic patterns would enable system administrators to determine, or "predict" where the traffic was heading. Once a known pattern is recognized, management would be able to mitigate the risk of attack by implementing additional defenses in that specific area.

Internal access to resources is another potential issue of security. Research has shown that by understanding the type of data stored and the level of authorized access held by an internal user, along with the frequency of access, an attacker can determine what function that user has within an organization (Kuhlmann & Schimpf, 2003). For example, if a user continuously accesses financial data (type of data stored), at an access level that is high (level of authorized access), over a period of six months (frequency of access), the attacker could determine the user is a financial officer, or a manager of finance. The major purpose of "role mining" is to determine where a user falls within a set of business roles within an organization. A method to ensure all access and access attempts are legitimate, management could log and store the successful and unsuccessful access attempts within our theoretical information security data warehouse. Even with a brief history of internal access logs, the data warehouse could determine a pattern of invalid user access attempts. Similar to the IP traffic monitoring, management could then determine which internal resources had the highest number of unsuccessful access attempts to identify what types of forms, files, and data that users are after. Even if the user is not maliciously trying to access resources outside of his privileges, the history of user attempts data could point to future attempts or breaches by internal employees that do have malicious intent.

Another information security threat to organizations that share their data with other organizations is breach propagation. As companies and organizations begin to share their information and data, knowledge sharing becomes an "inter-organizational" endeavor, which can positively increase the level and depth of knowledge. However the increase in inter-organizational reliance and

information sharing can lead to an increase in breaches across many other partner organizations. While valuable data and knowledge is being shared, so too are the risks and vulnerabilities of a security breach (Soper, Demirkan, & Goul, 2007). Soper, Demirkan and Goul (2007) present a model that attempts to minimize the effects of a security breach by centralizing security within a single security hub that controls access. In the event of a breach, the security hub can immediately notify all other hubs, while the breach can be contained to a specific segment or domain of the network. To adapt the study's breach propagation defense, the data security hubs could be implemented as data marts that could directly feed into our theoretical information security data warehouse. The data warehouse would store data of all breaches that have occurred, when they occurred, how they occurred, and any relevant circumstances surrounding the breach. This information could then be mined to determine if specific patterns of breach occurrences exist.

Another area that has recently been researched is how to analyze and control telecommunications frauds. Telecommunications are essential to networking because they are the pathway to how networks connect and communicate. A model for analysis and control of telecommunications fraud has six major components including detection, prevention, analysis, prediction, alarm, and control (Cao et al., 2004). Even though Cao et al. concluded prediction was an important element, it was never defined in the research.

The final adaptation of current research to discuss is centered upon finding bugs in software that are used to exploit security vulnerabilities for attacks. This method describes trends in ensuring software quality when systems are created by combining several different software components together to use specific features (Yin et al., 2007). This trend is known as service oriented architecture (SOA) and uses software as a service model. The studies implies that it is possible to use information on software bugs to predict the service security risks, or breach prediction at given levels of service. The model is split into three areas, software security, service composition, and hacking exposure (Yin et al., 2007).

## 5. PREDICTION MODEL

The prediction model being proposed will primarily use data collected on security breaches from multiple intra-organizational sources. The intra-organizational information security data warehouse would begin with a single organization collecting data from its various departments, divisions, regions, and business units. The benefit of collecting the security breach data across the different departments is a more comprehensive representation of the organization's security breach environment. There is a wealth of information available even within one department's experience with security breaches, but the data may not be as useful across an enterprise if it is not being shared, stored, and utilized by other departments. The input data should not be limited to just security breaches or attacks.

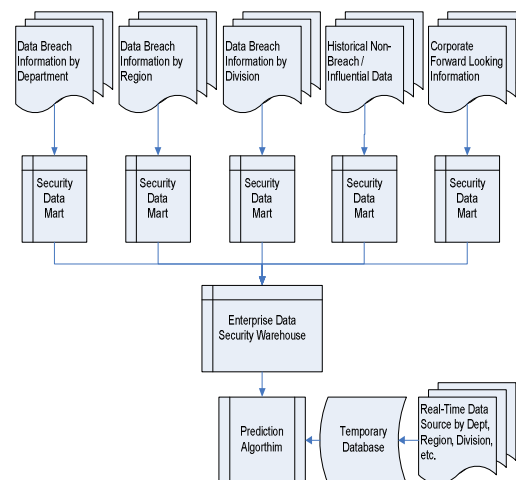
Data input into the security data warehouse must also include organizational information such as marketing data, public data releases, or human resource activity such as incentive plans and bonuses, or new hires, promotions, and terminations. The data should also include when a company is introducing a new product line or system upgrades because these types of data may provide a direct link or correlation to an increase in security attacks or breach attempts.

The incoming data can be obtained electronically from an IDS or an Intrusion Prevention System (IPS) in real time (Helmer et al., 1998; Lee et al., 2001). The data could be input into the data warehouse by a security administrator reviewing logs or by performing manual audits (Cao et al., 2004). The data could also be automatically fed by other systems like security hubs, data marts, and even other predictive models already in place (Kuhlmann & Schimpf, 2003; Soper, Demirkan, & Goul, 2007; Yin et al., 2007).

The multiple data feeds should be separated and scrubbed at the lowest level, and then combined and sent to a higher level for further scrubbing, and continued until the data reaches a level of meaningful summary (Helmer et al., 1998). The data will need to be organized and stored in a logical way that would allow the data warehouse administrator(s) to run multiple queries across various

data sources. The results of the queries can then be reviewed to determine if there is any relationship between security breach activities and other organizational activities. By collecting a brief history of security breach queries, an organization can examine these data and determine if an identifiable breach pattern emerges. The results should be used to inform management as to what data sources are most likely to be targeted and what, if any relationship exists between breach activity and intra-organizational happenings. If a significant pattern emerges, the organization may be able to not only defend against the attacks, but anticipate or predict when increased attacks may occur. This is an iterative and recursive process that repeats throughout the life of the organization. The model is shown in figure 1 below.

**Figure 1 – Prediction Model**



In this model, the data sources shown at the top come from different areas of the enterprise. This data is stored in some type of repository or data marts and is considered historical data of breaches and other relevant information. The forward-looking corporate data is information that has not been released but can have an effect on the security environment. These items include news release statements that have not been made public, potential purchases of different assets or other organizations, mergers, conversions to new technology, or any other type of information that can influence security. All of this information is input to a security data mart where the information is

scrubbed and transformed before being put into an enterprise data warehouse. Real-time data that is currently being streamed like IDS/IPS data, IP traffic data, news from other companies being released, and any other data will be stored into a temporary or real time database. All the information from the enterprise data warehouse and the temporary database will be input into an algorithm to determine what trends, or security events will likely occur. The temporary database is purged once data is stored long term or deemed no longer needed.

## 6. DISCUSSION

The security prediction model presented is from a broad perspective that eventually should be narrowed down and refined. The actual input parameters would not be known until an organization is able to identify and risk-rate what data is relevant and what is not. The actual algorithms and queries used to mine the data, summarize the data, and identify the breach patterns are also open. Even if one system used a specific set of inputs and algorithms, it does not mean a second system would be identical. The systems are likely to vary by organization and industry. Depending on what breach patterns eventually emerge, an organization may have to adjust its definition of what is valuable or high-risk data. And as organizations continually adapt to the ever-changing information landscape (Skovira, 2004), the prediction models will also evolve and mature.

The reason that news and other organizational data are important is because information thieves are becoming more opportunistic and some breaches occur because of corporate data or news releases (Campbell et al, 2003; Casey, 2006). A recent collection of information security breaches and the circumstances surrounding them reveals more than just technology security flaws (Panko, 2004). Many of the cases of information theft or breaches involve other factors that are related to the incident such as involuntary terminations, massive corporate layoffs or downsizing, espionage or theft of proprietary information by a competitor. Neohapsis (2002) reported on a Japanese aerospace company that experienced a security breach. This breach involved three employees at NEC and Toshiba Space Sys-

tems who were arrested for stealing data from Mitsubishi's antenna design for a high-speed Internet connection satellite. Police reported that one of the employees was able to gain unauthorized access to Mitsubishi's network that housed the proprietary data. It should be noted that it was widely publicized that NEC and Toshiba were competing against Mitsubishi to develop Internet technologies and for bidding on Japan's National Space Development Agency (NASDA) future projects. NASDA prohibited both NEC and Toshiba from bidding on NASDA projects for a month. In this example, a predictive security model may have equipped management to be more proactive in protecting its confidential data.

Even virus attacks can be targeted based on news or current events. A recent virus by the name of "Waledac.Trojan" was introduced on February 10, 2009, and was targeted to those sending electronic Valentine's Day greeting cards. According to Computer Associate's Security Advisor alerts, the Waledac Trojan has been observed to arrive in Valentine's Day-themed spam emails and spoofed websites. Organizations reported higher volume of network issues during Valentine's Day 2009 (Shanhaq, 2009). Scenarios like those mentioned above could have been stopped or mitigated by the prediction model if the right information and patterns were identified prior to the security breaches.

### Inter-Organizational Model

Ideally, this model would be more effective when multiple organizations are working together; creating an inter-organizational security data warehouse that crosses organizational lines and numerous industries. By having multiple sets of data from different areas, the identification of patterns and prediction should be more accurate and reliable. Organizations that reveal and share breach information may find some commonality between the reasons or circumstances revolving around the breach incident. Also, if one organization is experiencing a specific breach, and then other organizations from the same industry begin to experience similar breaches, the system could identify and alert the breaches in real time, giving other enterprises within the industry a warning of what is likely to come.

But a few issues present themselves with organizations sharing breach information with other companies. The biggest issue is that companies do not want to release information when breaches happen. Many companies will not publically release information on security breaches due to fear of threatening customer loyalty and negatively affecting the organization's reputation (Panko, 2004).

But even if the data was secure, preserving the privacy of the participating organizations is important as well. Privacy is an increasingly growing issue for integration and sharing of data (Clifton et al., 2007). Since multiple sources of data with varying degrees of proprietary and confidentiality will be coming into the data warehouse, the information must remain private. For example, if two companies who work in the defense industry as competitors decide to share a predictive data warehouse, since it would be beneficial to both, they need to establish practices that do not allow each other to query private data that is sensitive to their respective organizations (Zarsky, 2003).

Even though security and privacy issues need to be addressed, the benefits of a properly secured, shared information security predictive model can easily outweigh the risks. Because a shared model would not only share specific breach data, but also spread the burden of responsibility for preventing future attacks. A concentrated cooperative effort could also aid in tracking the origins of outbreaks like worms and help identify the source of the attacks.

## 7. CONCLUSION

As technologies advance, management must ensure that information security standards and practices also keep up to date. Although information security has usually been the responsibility of IT departments, some companies have made it a business issue as well as a technological one (Lohmyer, McCrory, & Pogreb, 2002). Most security defenses and countermeasures have traditionally taken a more passive approach to information security, primarily by installing firewalls, IDS systems, and authentication protocols. While these industry-standard

defense mechanisms serve as a foundational basis for securing information assets, we argue that organizations should start to consider initiating a more proactive and holistic approach to information security, specifically studying previous breach data, and identifying trends, in order to better anticipate or "predict" future security attacks or breaches before they occur.

Our proposed information security data warehouse model will be able to assist organizations in preparing for anticipated information security attacks or at a minimum to be more aware of attacks or breaches that have occurred in other companies and industries. Management could also utilize the data warehouse to determine its allocation of security resources, by providing more security in one area based on the risk and likelihood of an attack. Management could also identify practices and policies that are outdated. This is especially important since information security has a rapidly changing environment.

This paper presented a first attempt at developing a security prediction model that utilizes a data warehouse to store attacks and breaches on an organization. The reader is encouraged to critique, adapt, and modify the model to their specific organization or industry.

Future work will consist of blind reviews by professionals and researchers in evaluating and critiquing the model. The work will then continue in designing the information security data warehouse and archiving security breach information from several intra- and inter-organizational units. The data will be analyzed manually and data mining algorithms will be created to test the theory that breach patterns can be identified, and that the resulting future breaches can be predicted.

## 8. REFERENCES

- Agosta, L. (2000). *The essential guide to data warehousing*. Upper Saddle River, NJ: Prentice Hall.
- Boisot, M. (1998). *Knowledge assets: Securing competitive advantage in the information economy*. Oxford: Oxford University Press.

- Caceres, R., Duffield, N., Feldmann, A., Friedmann, J. D., Greenberg, A., Greer, R., Johnson, T., Kalmanek, C. R., Krishnamurthy, B., Lavelle, D., Mishra, P. P., Rexford, J., Ramakrishnan, K. K., True, F. D., & Merwe, J. (2000, May). Measurement and analysis of IP network usage and behavior. *IEEE Communications Magazine*, 38,144-151.
- Campbell, K., Gordon, L. A., Loeb, M. P., and Zhou, L. (2003). The economic cost of publicly announced information security breaches: empirical evidence from the stock market. *Journal of Computer Security*, 11 (2003), 431-448.
- Cao, L., Luo, C., Luo, D., & Zhang, C. (2004). Hybrid strategy of analysis and control of telecommunications frauds. *Proceedings of the 2nd international conference on information technology for application*, Harbin, China.
- Casey, E. (2006). Investigating sophisticated security breaches. *Communications of the ACM*, 49(2), February 2006, 48-54.
- Clifton, C., Doan, A., Elmagarid, A., Kantarcioglu, M., Schadow, G., Suciu, D., & Vaidya, J. (2004). Privacy-preserving data integration and sharing. *Proceedings of the 9<sup>th</sup> ACM SIGMOD workshop on research issues in data mining and knowledge discovery*, 19-26, Paris, France.
- Clifton, C., Jiang, W., Murugesan, M., & Nergiz, M. E. (2007). Is privacy still an issue for data mining? *National science foundation symposium on next generation of data mining and cyber enabled discovery for innovation*, Baltimore, USA.
- Cole, E., Krutz, R., & Conley, J. W. (2005). *Network security bible*. Indianapolis, IN: Wiley Publishing.
- CSI Computer Security Institute. (2008). *Computer crime and security survey*. San Francisco, CA: Computer Security Institute. Retrieved from <http://www.gocsi.com>.
- Gallegos, F., Senft, S., Manson, D. P., & Gonzales, C. (2004). *Information technology control and audit* (2<sup>nd</sup> Ed.). New York: Auerbach Publishing.
- Helmer, G. G., Wong, J. S. K., Honavar, V., & Miller, L. (1998). Intelligent agents for intrusion detection, *Proceedings of Information Technology Conference*, 121-124, New York, USA.
- Kuhlmann, M., & Schimpf, G. (2003). Role mining: Revealing business roles for security administration using data mining technology, *Proceedings of the eighth ACM symposium on access control models and technologies*, 179-186, Como, Italy.
- Lee, W., Stolfo, S. J., Chan, P. K., Eskin, E., Fan, W., Miller, M., Hershkop, S., & Zhang, J. (2001). Real time data mining-based intrusion detection, *Proceedings of DARPA information survivability conference and exposition II (DISCEX'01)*, Vol(1), 89-100, Anaheim, CA.
- Lohmyer, D.F., McCrory, J., and Pogreb, S. (2002). Managing information security. *The McKinsey Quarterly 2002 Special Edition: Risk and Resilience*, 12-15.
- Mitnick, K. D. and Simon, W. L. (2002). *The art of deception: Controlling the human element of security*. New York: John Wiley & Sons.
- Neohapsis. (2002). Aerospace workers arrested for hacking. Retrieved from <http://archives.neohapsis.com/archives/sn/2002-q2/0291.html>
- Panko, R. R. (2004). *Corporate computer and network security*. Upper Saddle River, NJ: Pearson Education.
- Ranum, M. J. (1993). Thinking about firewalls, *Proceedings of second international conference on systems and network security and management (SANS-II)*, April, 1993.
- Samuelle, T. J. (2008). *Mike Meyers' certification passport: CompTIA Security+*. New York: McGraw Hill.



- Shanbhaq, R. (2009). CA Issues Early Warning of Possible Waledac Trojan on Valentine's Day. Retrieved from <http://sip-trunking.tmcnet.com/topics/security/articles/50095-ca-issues-early-warning-possible-waledac-trojan-valentines.htm>.
- Simon, A. R. (1997). *Data warehousing for dummies*. Hoboken, NJ: Wiley Publishing.
- Skovira, R. J. (2004). Using informational landscape as a model to understand information use and design within organizations. *Issues of Information Systems*, 5(1), 308-314.
- Soper, D. S., Demirkan, H., & Goul M. (2007). An interorganizational knowledge-sharing security model with breach propagation detection. *Information Systems Frontier*, 9(5), 469-479.
- White, G., & Conklin, W. M. A. (2008). *All in one CompTIA security+ exam guide (2<sup>nd</sup> Ed)*. New York: McGraw Hill.
- Yin, J., Tang, C., Zhang, X., & McIntosh, M. (2007). On estimating the security risks of composite software services. International Business Machines (IBM). Retrieved from <http://research.ihost.com/password/papers/Yin.pdf>.
- Zarsky, T. Z. (2003). "Mine your own business!": Making the case for the implications of data mining of personal information in the forum of public opinion. *Yale Journal of Law and Technology*, 5, 1-57.