https://doi.org/10.48009/4_iis_2016_51-57

Issues in Information Systems
Volume 17, Issue IV, pp. 51-57, 2016

WHO'S KNOCKING AT MY DOOR: DEVELOPING A WEB SERVICE VISUALIZATION TOOL FOR MONITORING USER ACCOUNT HEALTH

Thomas A. Chapman, University of Mississippi, tchapman@bus.olemiss.edu

ABSTRACT

Identifying potential threats to network security is arguably the single most important task for today's IT professional. This paper seeks to aid in that task by outlining a framework whereby a readily available web-service visualization tool can be developed. Such a tool would accurately visualize one important aspect of an SNMP notifications log. This paper describes a thought experiment, which outlines the conceptual steps necessary to produce and deploy a web-service visualization tool that IT administrators could use to determine if individual user accounts are at risk of being compromised. Conceptually, the web-service would run on top of a predictive model that exists within the popular data analytics software, RapidMiner. This predictive model would take a text extraction of an SNMP notification log as its input and produce a digital map of user accounts that may be at risk of being compromised.

Keywords : Cybersecurity, SNMP Notifications, Log Analysis, Data Visualization

INTRODUCTION

The financial cost of recent cyberattacks in the business industry is staggering. However, it does not reflect the sum-total cost of cyber-attacks. While news of significant data breaches against corporate and financial targets fill the national headlines on a seemingly daily basis, similar attacks against colleges and universities are less widely reported. According to a study by the Ponemon Institute, cyber-attacks against institutions of higher learning accounted for around nine percent of the total number of cyber-attacks that were reported in 2014. The total cost associated with those attacks was approximately 3.2 million dollars (Ponemon Institute, L.L.C., 2014).

Within universities, individual departments often have only a few individuals, perhaps one or two IT professionals, who work to handle all the normal job functions incumbent in larger, centralized IT departments. The ability of such individuals to do a detailed security analysis of individual LANs may be limited due to the amount of time that an IT administrator who has many different demands on their time may devote to such a task.

In order to define why such analysis is necessary, it is first necessary to determine the extent to which individual department LANs at colleges and universities are experiencing some form of cyber-attack. The guiding principle for what constitutes an attack has consistently been any action that is undertaken with malicious intent, which affects one of the three basic computer security principles of availability, integrity, or confidentiality (Myerson, 2002). In order to assess whether they are under threat, IT administrators have traditionally relied on data from two disparate, but complementary, sources.

The first source of data comes from Simple Network Management Protocol (SNMP) notifications that are thrown off by the network switches. The second type of data typically comes from some class of network monitoring device and/or software package. Within this general classification, Intrusion Detection Systems (IDS) are a subclass, which remain popular within the IT security community (Corchado & Herrero, 2011).

Within the context of a computer network, an intrusion detection system is a tool that can detect suspicious patterns in a network packet flow. The proposed architecture for the IDS concept was first articulated in 1987 when it was hypothesized that the exploitation of a computer system's vulnerabilities involved abnormal use of the system. (Denning, 1987) Examples of anomalies that can be detected, which Denning cites in her paper, include an

abnormally high number of password failures for a single user, abnormal login times, and unexpected login times from off-site locations.

However, intrusion detection systems are expensive and sometimes difficult to configure. As such, they may not represent a viable option for IT administrators who are working at the individual department level at large universities, or at colleges with small operating budgets. According to a recent survey by SC magazine, which is geared towards security professionals, the cost of a commercial intrusion detection system can run in the many thousands of dollars (SC Magazine, 2016). While there are free IDS solutions available, they can be difficult to find and/or configure for administrators who have little time or money and who may not be well versed in either the need for an IDS solution, or the capabilities of the various solutions that are available.

In addition to the cost to purchase and configure an IDS solution, the absence of agreed upon standards of best practice for identifying and diagnosing traffic anomalies continues to be a hindrance towards the development of systematized cyber security controls (Barford, et. al., 2002). Largely, network administrators rely on their own experience to generate ad-hoc rules, which may vary significantly depending on the context in which they're applied.

Clearly, a free alternative to expensive commercial intrusion detection systems is required. To that end, the RapidMiner data analysis software package provides a viable alternative for analyzing such data. RapidMiner was developed as a way to perform a multitude of analytical techniques on many different types of data. It is a software platform package that provides an integrated environment for performing a variety of machine learning, data mining, text mining, predictive analytics and business analytics. It was first developed by three researchers out of the Artificial Intelligence Unit of the Technical University of Dortmund as YALE (Yet Another Learning Environment) in 2001 (Deutsch, 2010). In 2007, with the commercial success of the product, the name was changed from YALE to RapidMiner. In 2016, RapidMiner was recognized by Gartner Research Group as the leader in advanced analytics for the third year in a row (RapidMiner, 2016). Despite its success, however, RapidMiner has remained true to its open source roots. It retains both a free version of its popular RapidMiner Studio software as well as an academic version, which it licenses free to researchers who are affiliated with a public, educational institutions.

It should be possible, therefore, to develop a statistical model, which will run within the RapidMiner Studio software package, that can accept a text extracted .csv file of an SNMP notification log as input and produce a predictive model of which user accounts are at risk, based on a baseline set of normative values established previously.

RELATED WORK

Innovative data visualization techniques within the cyber security field have garnered increasing attention from researchers in recent years as the frequency and severity of cyber-attacks has continued to rise.

Giacobe and Xu, from the Pennsylvania State University, first looked at applying geographic information systems visualization techniques to the cyber security field in 2011 as part of the Visual Analytics Science and Technology challenge that was held as part of that year's IEEE symposium (Giacobe & Xu, 2011). In order to realize their visualization technique, the researchers used the GeoViz toolkit along with a mapping software, ArcGIS, to construct a simulated network "landscape" comprised of servers, workstations, and the Internet. The authors found that they were able to view data anomalies in an integrated fashion across multiple views by using the GeoViz toolkit.

Ferebee, et. al., looked at additional visualization techniques for cyber security data in 2011 (Ferebee, et. al., 2011). They applied meteorology-based storm mapping techniques to cyber security data. The following year, Zhao, et. al followed up their 2011 GIS visualization technique with a 2012 federated approach in which they augmented the geographic tools of their previous paper with a custom-designed frequency analysis plotting technique that they developed in the Processing program language (Zhao, et al., 2012).

In addition to developing innovative visualization techniques for cyber security data such as those listed above, numerous researchers have looked at ways to classify anomalous network signals as either malicious or benign.

Barford, et al, for example, used wavelet filters to expose the details between ambient and anomalous network traffic (Barford, et. al., 2002). The authors collected network IP flow data and SNMP measurements from one of the border routers at the University of Wisconsin and then worked to separate possible anomalies into two groups. The first group that they analyzed consisted of flash crowd events, which typically lasted a week or more in duration. The second group of the events that they analyzed were short-lived events and consisted of everything from network failures to actual intrusion attempts. In order to test that anomalies within this second group of events represented real cyber-attacks, the authors developed a method of exposing short term, high intensity events, which was based on computing the normalized local variance of the mid and high frequency parts of the network packet flow. In addition, such events were later validated as intrusion attempts by IT security engineers using more traditional intrusion detection methods.

Other authors who have studied network fault detection in a similar manner include Hood and Ji, who proposed an early system of fault detection in 1997 that was based on adaptive statistical techniques formed around a Bayesian framework (Hood & Ji, 1997). According to the authors, their approach allowed them to detect unknown faults, correlate information in space and time, and detect subtle changes occurring before the actual failure. Hood and Ji's work complements the work of Katzela and Schwartz, who identified similar schemes for fault identification in communication networks (Katzela & Schwartz, 1995).

The methods described in the papers above rely on the extraction of pertinent information from a network data medium such as an SNMP query log or network packet flow obtained from an IDS or some such similar device. Classified generally, this is the essence of modern Business Intelligence applied to network log data. Chen, Chiang, and Storey provide a nice synopsis of BI applications and techniques in their 2012 paper (Chen, Chiang, & Storey, 2012). Specifically, techniques of interest for the present research appropriately fall within the classification of log analysis. Researchers who have studied log analysis in relationship to web query logs include Beitzel, Jensen, Chowdhury, Grossman & Frieder (Beitzel, et. al., 2004). Later, Jansen examined search logs from a log analysis perspective in which he provided a solid overview of log analysis techniques (Jansen, 2006). Several researchers including Jon Stearley, John Rouillard, Paul Barford, and others have used log analysis on SNMP notification logs (Stearley, 2004; Rouillard, 2004; Barford, et. al., 2002).

CONCEPTUAL MODEL

This research centers on the question of whether IT administrators who work at the department level at large universities or at colleges with small operating IT budgets, and who may or may not be trained in IT security but who have responsibility over security for one or more LANs, routinely analyze the SNMP notification logs from their network switches.

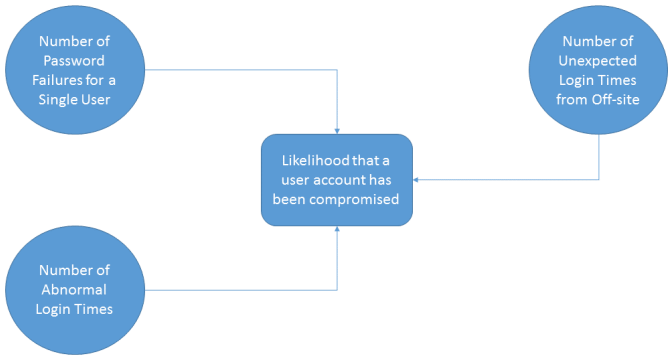

Figure 1. Conceptual Model of Factors Involved in Determining the Likelihood that a Target Computer has been Compromised

Independent Variables:

- IV1 – The number of password failures for a single user
- IV2 – The number of abnormal login times within a given time period
- IV3 – The number of unexpected login times from offsite locations

The research hypotheses to be tested, therefore, are as follows:

P1 – The likelihood that a user account has been compromised will be positively related to the number of password failures for a single user.

P2 – The likelihood that a user account has been compromised will be positively related to the number of abnormal login times within a given time period.

P3 – The likelihood that a user account has been compromised will be positively related to the number of unexpected login times from offsite locations.

METHOD

Each type of network hardware switch has its own SNMP protocols that it uses. Two popular examples are Cisco and Hewlett-Packard, both of which are industry leaders in the network infrastructure market. By default, both types of switches are enabled to send SNMP notifications. Hewlett-Packard switches, for example, will throw an SNMP notification, which is structured with an Event ID, a Method, an IP Address Type, an IP Address, a User Name, and a date and time (Hewlett-Packard, 2016). Cisco switches structure their SNMP notifications in a slightly different manner. However, the basic data that the SNMP notification log contains is the same (Cisco, 2016). One possible Event ID that can be included in both vendors' SNMP notifications indicates an invalid password that is entered through either a direct serial, Telnet, or SSH connection.

As such, it is possible to use established log analysis techniques to extract abnormal login events from the SNMP notification log and use them to populate a comma-separated value (.csv) file. RapidMiner, in turn, can accept such a file as a data input source (RapidMiner, 2016).

In order to test hypothesis 1, the number of invalid login attempts per user account will be analyzed and those which show a statistically significant level of invalid attempts from a normative standard will be flagged. Similarly, the date and time field can be extracted and analyzed separately to test hypothesis 2, or in combination with the IP address field to test hypothesis 3.

Two possible avenues for establishing a normative standard exist. First, it should be possible to establish baseline values based on a historical sampling of SNMP logs taken at different points in the lifecycle of the LAN to be queried. This method relies on the availability of historical SNMP log data. If such data exists, however, a time series model could be developed, which tracks abnormal login events per user account based on the three propositions listed above and then determines if the current values are statistically different from the historical norm. If so, that would indicate that the user account in question has been, or is in danger of being, compromised.

A second option for establishing a baseline normative standard would involve taking a random sampling of the SNMP logs from department LANs across numerous universities and colleges in an effort to establish a set of population parameters. However, this would require individual IT administrators to submit their SNMP logs to the researcher for text extraction. Such extraction would obviously incorporate a process that would automatically de-identify specific network information such as IP addresses. However, trying to convince IT administrators to participate in such research could prove to be difficult, to say the least.

Nonetheless, this second option is not dissimilar from how the final web service would operate. Theoretically, the web service would allow the user to import their SNMP notification log, so that the relevant data could be extracted into a .csv file, which would serve as the useable input file for the RapidMiner statistical model. User accounts that show a high dimensionality in two or more of the independent variables would be clustered together and displayed graphically so that the administrator can easily visualize which user accounts may be at risk of being compromised.

SUMMARY

This paper describes a thought experiment which outlines some of the conceptual steps that are possible for the development of a web-service visualization tool, which IT administrators could use to quickly visualize one important aspect of their SNMP notification logs.

Log analysis is an important tool in any IT administrator's arsenal to protect against possible cyber-attacks. However, IT administrators who have multiple demands on their time may not be able to put forth the time necessary to routinely do systematic log analysis in search of abnormal events.

The techniques described in this research explore the possibility of developing a data visualization tool that runs on top of a statistical model running within the popular data analytics software, RapidMiner. The RapidMiner Studio product started as a research tool within academia. It evolved into a commercial product that has consistently been recognized by the Gartner Group as a leading product in the data analytics market. Even though it is a commercial product, it retains its open-source roots through several licensing options that academic researchers can use.

The conceptual model described herein relies largely on Denning's 1987 work on intrusion detection systems. In that paper, Denning theorizes that data intrusion patterns can be detected as abnormalities in the normal operating patterns of network systems. Intrusion detection systems, therefore, identify such abnormalities. In doing so, they have been widely adopted by IT administrators as a primary line of defense against the possibility of a cyber-attack.

However, commercial intrusion detection systems can often be expensive and difficult to configure. Acquiring and configuring the right IDS may be a challenge for IT administrators working at the individual department level at large universities, or at colleges with small operating budgets. Therefore, it could be highly useful for such administrators to have a free web-service tool available to them, which can perform one important function of a commercial IDS.

One of the biggest obstacles to establishing such a tool, however, is to establish a normative standard of user account data for individual LANs, against which to compare current event log abnormalities. Two possible avenues for establishing such a standard exist. The first is to develop a historical standard for the individual LAN to be queried. However, this method depends on the existence of historical SNMP log data, which may or may not be available. The second method is to establish a normative standard based on a random sampling of individual LAN's SNMP log data from several universities and colleges. This method is potentially problematic from a data collection standpoint as many IT administrators could be hesitant about handing over their network logs to an academic researcher. Even though controls would be put into place in order to fully de-identify data, IT administrators are understandably guarded about sharing such data.

If such obstacles can be overcome, however, the potential benefits could prove to be a significant improvement in cybersecurity preparedness for institutions of higher learning.

REFERENCES

Barford, P., Kline, J., Plonka, D., & Ron, D. P. (2002). A Signal Analysis of Network Log Anomalies. *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet Measurement*, pp. 71-82.

Beitzel, S. M., Jensen, E. C., Chowdhury, A., Grossman, D., & Frieder, O. (2004, July). Hourly analysis of a very large topically categorized web query log. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 321-328). ACM.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *Mis Quarterly*, 1165-1188.

Cisco. (2016, May 13). *Cisco Network Switch Configuration*. Retrieved from Cisco.com: http://www.cisco.com/c/en/us/td/docs/optical/15000r9_2/dwdm/troubleshooting/guide/b_454d92_ts/b_454d92_ts_chapter_011.pdf

Corchado, E., & Herrero, A. (2011). Neural Visualization of Network Traffic Data for Intrusion Detection. *Applied Soft Computing, 11*(2), 2042-2056.

Denning, D. (1987). An Intrusion-Detection Model. *Software Engineering, IEEE Transactions on Software Engineering, 13*(2), 222-232.

Deutsch, G. (2010, March 18). *RapidMiner from Rapid-I at CeBIT.* Retrieved from Data Mining Blog: http://www.data-mining-blog.com/cloud-mining/rapidminer-cebit-2010/

Ferebee, D., Dasgupta, D., Schmidt, M., & Wu, Q. (2011). Security Visualization: Cyber Security Storm Map and Event Correlation. *IEEE Symposium on Computation Intelligence in Cyber Security* (pp. 171-178). Paris: IEEE.

Giacobe, N. A., & Xu, S. (2011). Geovisual Analytics for Cyber Security: Adopting the GeoViz Toolkit. *Visual Analytics Science and Technology*, (pp. 313-314).

Hall, D., & McMullen, S. (2004). *Mathematical Techniques in Multisensor Data Fusion.* Boston, MA: Artech House.

Hewlett-Packard. (2016, May 13). *Configuring SNMP*. Retrieved from Hewlett-Packard Enterprise: http://h22208.www2.hpe.com/eginfolib/networking/docs/switches/K-KA-KB/15-18/5998-8160_ssw_mcg/content/ch06s10.html

Hood, C., & Ji, C. (1997). Proactive Network-Fault Detection. *IEEE Transactions on Reliability*, 333-341.

Jansen, B. J. (2006). Search log analysis: What it is, what's been done, how to do it. *Library & Information Science Research, 28*(3), 407-432.

Katzela, I., & Schwartz, M. (1995). Schemes for Fault Identification in Communication Networks. *IEEE/ACM Transactions on Networking*, 753-764.

Myerson, J. (2002). Identifying Enterprise Network Vulnerabilities. *International Journal of Network Management*(12), 135-144.

Ponemon Institute, L.L.C. (2014). *2014 Cost of Cyber Crime Study.* HP Enterprise Security.

RapidMiner. (2016, May 13). *Gartner Magic Quadrant for Advanced Analytics Platforms*. Retrieved from RapidMiner.Com: https://rapidminer.com/resource/leader-gartner-magic-quadrant-advanced-analytics/

RapidMiner. (2016, May 13). *RapidMiner Documentation*. Retrieved from RapidMiner: http://docs.rapidminer.com/studio/how-to/

Rouillard, J. P. (2004, November). Real-time Log File Analysis Using the Simple Event Correlator (SEC). *LISA, 4*, pp. 133-150.

SC Magazine. (2016, May 13). *Intrusion Detection System Reviews*. Retrieved from SC Magazine: http://www.scmagazine.com/intrusion-detection-systems/products/91/0/

Stearley, J. (2004, September). Towards informatic analysis of syslogs. InCluster Computing, 2004 IEEE International Conference on (pp. 309-318). IEEE.

Ten, C. W., Liu, C. C., & Manimaran, G. (2008). Vulnerability Assessment of Cybersecurity for SCADA Systems. *IEEE Transactions on Power Systems*, 1836-1846.

Zhao, M., Zong, C., Ciamaichelo, R., Konek, M., Sawant, N., & Giacobe, a. N. (2012). Federating GeoVisual Analytic Tools for Cyber Security Analysis. *IEEE Symposium on Visual Analytics Science and Technology* (pp. 303-304). Seattle, WA: IEEE.