Presented at the DLSU Research Congress 2015
De La Salle University, Manila, Philippines
March 2-4, 2015

Network Analysis with Report Authoring (NARA)

Reginald Caldez[1], Jan Gilbert Castro[2], Kelvin Richie Co[3], and Geanne Ross Franco[4]
[1,2,3,4]Computer Technology Department, De La Salle University, 2401 Taft Ave., Manila

Abstract: People use their smartphones, tablets, laptops, or a combination of these to connect to the internet via the organization's network. These devices are used to access various websites. Thus, organizations implement policies that limit access to only work-related websites. However, these policies may be bypassed and users are able to access sites that should be blocked. This study aims to utilize a report authoring tool to make a set of templates, used for the generation of reports. The produced reports assist the organization in identifying websites that may have bypassed these policies. This study helps administrators in making changes to the policies that have been implemented. The study collects two sets of data, specifically the before and after the application of changes to the policies. The reports produced help the users determine if the changes implemented to the network served its function.

The group has collected network data from a computer laboratory of about twenty hosts. The group provides a preliminary list of frequented websites based of a sample set of data. The group's findings can be used to see the most frequented visited websites by the students. By using the initial data collected, the system administrator may be able to identify if the websites being accessed should be allowed or blocked.

Key Words: Content Filter; Packet Sniffing; Data Aggregation; Report Authoring

1. INTRODUCTION

The internet is essential; almost everyone uses it to perform their daily tasks. In an organization, multiple users and devices are connected to its network which allows access to the internet. Thus securing this network is of utmost importance to these organizations in order to protect their assets.

Security measures such as content filters are therefore implemented in these organizations in order to make the network more secure. Content filters block the user's access to websites which may contain malicious content and websites which are unrelated to work. However, content filters may be bypassed because of various reasons such as outdated website databases or incorrect website categorization.

The effectiveness of the content filters is now in question and must be monitored through packet sniffers such as Wireshark. Monitoring an entire network can be done through packet sniffing the mirrored port of the default gateway of the network. Port mirroring is achieved using at least two ports within a switch, wherein one specifies a source port and a destination port. The source port is used as the source of the data. The data from the source port is

copied or "mirrored" then sent to the assigned destination port and its original destination. The collected data from the packet sniffer are stored within a database.

The data collected are used for data analysis. Data analysis examines the available data and draws conclusions from them. In this study, data analysis allows the system to show the user which websites are bypassing the content filter implemented. The collected data from the network are categorized based on the content of each websites. Certain categories such as games, entertainment, sports, and social media are restricted by the content filter. Therefore, one can assume that users who have accessed sites within these categories have bypassed the content filter.

Report authoring enables users to present the analyzed data through informative and visual reports. These reports may contain various objects such as tables, graphs and charts which can help users get a better understanding of the data. One example of a report authoring tool is the Report Designer within Microsoft's Visual Studio. It is used to define the contents of a report and can be used with other Microsoft products. IBM Cognos 8 Business Intelligence suite is another existing solution for business companies. This suite allows it users to view or create business reports and analyze data.

While there are existing business intelligence solutions, the same cannot be said for network data. The complexity of network data cannot be easily represented using basic reports. Reports generated by current network tools show only raw data.

The goal of this study is to create a system that assists in the identification of websites which may have bypassed the content filter in place. This is achieved by integrating the collected network data with an existing report authoring tool such as Microsoft's Report Designer, or IBM Cognos 8. Through the use of report authoring tools, one is able to create reports which can assist in identifying wrongly categorized or uncategorized websites, which is essential in identifying websites which have bypassed the content filter in place.

The reports generated can provide the user a general outline of how the network resources are being utilized, which sites are being accessed, the categories of these sites, the time at which these are accessed, and other essential information.

2. METHODOLOGY

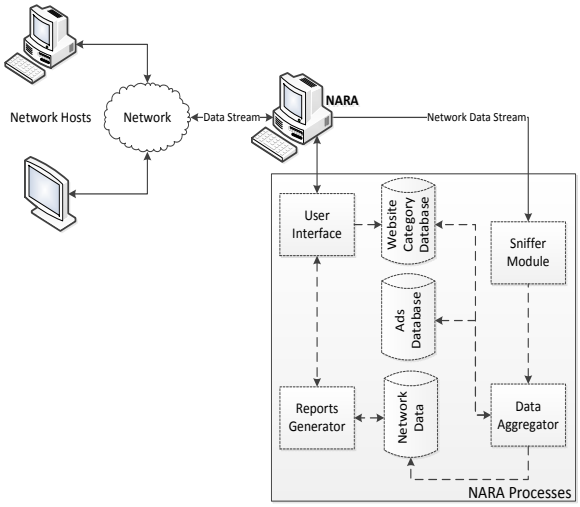

Fig. 1. NARA System Overview

NARA serves four main functions: data collection, data aggregation, report generation, and report viewing. This is implemented using the setup presented in Figure 1. Wherein, NARA is connected to an existing network in order to collect the incoming and outgoing data generated by its users.

NARA connects to a switch port that is configured for port mirroring. With this, all network traffic passing through the network's default gateway are mirrored and sent to NARA. NARA's packet sniffer collects the HTTP request, HTTP response, and HTTPS traffic

The collected data is sent to a data aggregator; the data aggregator performs TCP reconstruction and creates a DNS table. This is done to get the status-code of the HTTP traffic and the domain of the HTTPS traffic. Each entry from the HTTP and HTTPS traffic are classified into their specific categories. These categories are used to determine if a certain website is restricted or allowed by the content filter because certain categories are labeled as restricted.

The categorized entries are stored into a database which is connected to the report generator. The report generator utilizes a report authoring tool and the collected categorized data, in order to provide reports of the network data.

A report authoring tool provides method to design and create templates for the reports. NARA will provide templates for the user to select. Once a template has been selected, this is relayed to the report authoring tool and a report is generated with the necessary data. The generated report is sent to the user interface for viewing.

3. RESULTS AND DISCUSSION

Table 1. List of Hostnames Requested on October 9, 2014

Timestamp	Hostname	Status Code
8:37 AM	www.google.com	HTTP/1.1 302 Found
8:37 AM	netacad.com	HTTP/1.1 301 Moved Permanently
9:11 AM	www.food.com	HTTP/1.1 200 OK
9:11 AM	www.foodonthetable.com	HTTP/1.1 200 OK
9:12 AM	www.foodnetwork.com	HTTP/1.1 403 Forbidden
9:12 AM	www.scrippscontroller.com	HTTP/1.1 200 OK
9:12 AM	mail.yahoo.com	HTTP/1.1 302 Found
9:17 AM	www.washingtonpost.com	HTTP/1.1 200 OK
9:17 AM	www.twitter.com	HTTP/1.1 304 Not Modified
9:17 AM	www.facebook.com	HTTP/1.1 403 Forbidden
9:17 AM	www.burstnet.com	HTTP/1.1 200 OK
9:18 AM	en.wikipedia.org	HTTP/1.1 200 OK
9:18 AM	www.who.int	HTTP/1.1 200 OK
9:21 AM	www.theguardian.com	HTTP/1.1 200 OK
9:24 AM	lifestyle.inquirer.net	HTTP/1.1 302 Found
9:24 AM	maps.google.com	HTTP/1.1 200 OK
9:24 AM	newsinfo.inquirer.net	HTTP/1.1 200 OK
9:24 AM	business.inquirer.net	HTTP/1.1 200 OK
9:24 AM	www.seattlepi.com	HTTP/1.1 200 OK
9:24 AM	www.buzzfeed.com	HTTP/1.1 403 Forbidden
9:25 AM	log.pinterest.com	HTTP/1.1 200 OK

The data shown in Table 1 is the partial entries of the data collected on October 9, 2014. The data was collected from the existing setup of Center for Network and Information Security laboratory which uses the content filter Sophos. The CNIS laboratory has about twenty workstations including the workstation used by the instructor. The total amount of entries collected on October 9, 2014 reached twenty thousand entries. These entries were generated using a part of the data aggregator module that reconstructs each TCP's requests with its corresponding responses.

The data within the table shows only the unique hostnames; this is done by getting the first entry of each unique hostname in the collected data. The table shows the timestamp of each request, the hostname of the website being accessed, and its status-code.

Table 1. Legend for Status Codes in Table 1

Status Code	Meaning
200 OK	Allowed by the Firewall
301 Moved Permanently	Response to requests where the URL is redirected permanently
302 Found/Moved Temporarily	Response to requests where the URL is redirected temporarily
403 Forbidden	Valid Request. Blocked by the Firewall

The status-code is the component that determines if the website was allowed or blocked by Sophos. The entries have not been categorized yet so it is still difficult to assess if a particular website has bypassed Sophos or not.

Table 3. Table of Allowed and Blocked Sites by Sophos

Restriction Status	No. of Requests	Percentage
Allowed	16875	83.66%
Blocked	3296	16.34%
Total	20171	

The data shown in Table 3 shows the number of entries which were allowed and blocked by Sophos on October 9, 2014. As seen on Table 3, Sophos allowed more entries which could mean that on Oct. 9, 2014, students were accessing sites that were within categories allowed by Sophos.

Table 4. List of Hostnames Requested on November 10, 2014

Timestamp	Hostname	Status Code
7:24 AM	www.microsoft.com	HTTP/1.1 304 Not Modified
8:32 AM	netacad.com	HTTP/1.1 301 Moved Permanently
8:33 AM	ccs1.dlsu.edu.ph	HTTP/1.1 302 Redirect
8:34 AM	mail.yahoo.com	HTTP/1.1 302 Found
8:34 AM	go.microsoft.com	HTTP/1.1 302 Found
8:40 AM	www.facebook.com	HTTP/1.1 403 Forbidden
8:44 AM	ask.fm	HTTP/1.1 200 OK
9:12 AM	www.google.com	HTTP/1.1 302 Found
9:14 AM	platform.twitter.com	HTTP/1.1 304 Not Modified
9:23 AM	pcx.com.ph	HTTP/1.1 200 OK
9:25 AM	www.laptopmag.com	HTTP/1.1 200 OK
9:25 AM	www.toptenreviews.com	HTTP/1.1 200 OK
9:31 AM	www.youtube.com	HTTP/1.1 403 Forbidden
9:31 AM	en.wikipedia.org	HTTP/1.1 200 OK
9:33 AM	www.cdrking.com	HTTP/1.1 200 OK
9:41 AM	www.wikihow.com	HTTP/1.1 200 OK
9:41 AM	www.azcentral.com	HTTP/1.1 302 Moved Temporarily
9:42 AM	www.linkedin.com	HTTP/1.1 200 OK
9:42 AM	log.pinterest.com	HTTP/1.1 200 OK
9:42 AM	www.buzzfeed.com	HTTP/1.1 403 Forbidden
9:55 AM	instagram.com	HTTP/1.1 200 OK

The data shown in Table 4, is the partial entries of the data collected on November 10, 2014. The setup for the data which was collected is the same as the data collection on October 9, 2014. The total amount of entries collected on November 10 was about fifteen thousand entries. The data collected from October 9 and November 10 had some similar sites which were visited. As seen in tables 1 and 4, some of these are: netacad.com, mail.yahoo.com, www.facebook.com, www.google.com, en.wikipedia.org, log.pinterest.com, and www.buzzfeed.com. The data gathered from these sites particularly the status-code were consistent between the two tables, even though the time and dates differ. Therefore, it is safe to assume that the content filter is functioning and the only thing still in question is if it is correctly allowing and blocking the websites being accessed by the users.

4. CONCLUSIONS

In conclusion, an organization must ensure the security of its network and its assets. Thus, security measures like content filters must be implemented. However, these content filters may be bypassed because of several reasons. And because of this, the effectiveness of these content filters are now in question. In order to ensure the effectiveness of the content filter implemented, an organization must collect the data of its entire network that is done through the implementation of a port mirror and mirroring the network traffic passing. The collected data must be categorized and analyzed to determine if certain websites have been accessed even though they should have been restricted.

From the current available data collected, the group has assumed that the content filter is indeed functioning. The group has come to this conclusion because of the consistency of the responses or status-codes through differing dates.

In order to achieve the goal of this study, the group must still collect additional data for further analysis, categorize the data that have been collected, integrate the collected data to a report authoring tool and generate reports through the use of the report authoring tool.

5. ACKNOWLEDGMENTS

The authors wish to acknowledge and thank their adviser Ms. Geanne Franco, and their panelists for the support and guidance they have

given to the group.

6. REFERENCES

Aho, A. & Ullman, J. (1992). Chapter 10. "Patterns, Automata, and Regular Expressions". Foundations of Computer Science. Retrieved June 5, 2014, from http://infolab.stanford.edu/~ullman/focs/ch10.pdf

EasyDNS. (n.d.). What is Reverse DNS. Retrieved June 3, 2014, from http://helpwiki.easydns.com/index.php/What_is_Reverse_DNS

Hillyer, M. (n.d.). An Introduction To Database Normalization. Retrieved June 9, 2014, from http://mikehillyer.com/articles/an-introduction-to-database-normalization

IBM Corporation. (2011, April). Cognos 8 Business Intelligence: Getting Started. Retrieved October 25, 2013, from http://download.boulder.ibm.com/ibmdl/pub/software/data/cognos/documentation/docs/en/8.4.0/wig_cr.pdf

IBM Corporation. (2013). Report authoring with IBM Cognos Business Intelligence. Retrieved October 31, 2013, from http://public.dhe.ibm.com/common/ssi/ecm/en/ytw03198caen/YTW03198CAEN.pdf

Karttunen, L. and Zwicky A. (1985). Natural Language Parsing. Retrieved July 22, 2014, from http://web.stanford.edu/~zwicky/natural-language-parsing-intro.pdf

Lee, T. (1999). Information Modeling: From Design to Implementation. Retrieved July 24, 2014, from http://www.mel.nist.gov/msidlibrary/doc/tina99im.pdf

Microsoft. (n.d.). Description of the database normalization basics. Retrieved June 12, 2014, from http://support.microsoft.com/kb/283878

Microsoft. (n.d.). ODBC--Open Database Connectivity Overview. Retrieved July 24, 2014, from http://support.microsoft.com/kb/110093

Microsoft. (n.d.). Report Authoring. Retrieved October 31, 2013, from http://technet.microsoft.com/en-us/library/aa256343%28v=sql.80%29.aspx

Microsoft. (n.d.). What Is ODBC?. Retrieved June 12, 2014, from http://msdn.microsoft.com/en-us/library/windows/desktop/ms714591(v=vs.85).aspx

Microsoft. (n.d.). Writing SQL Queries: Let's Start with the Basics. Retrieved June 9, 2014, from http://technet.microsoft.com/en-us/library/bb264565(v=sql.90).aspx

Moore, A., & Zhang, J. (2007). Traffic Trace Artifacts due to Monitoring Via Port Mirroring. Retrieved October 25, 2013, from http://www.cl.cam.ac.uk/~awm22/publications/zhang2007traffic.pdf

NTCHosting. (n.d.). SQL (Structured Query Language). Retrieved July 22, 2014, from http://www.ntchosting.com/databases/structured-query-language.html

Petrunia, S. (2007, April). MySQL Optimizer and Prepared Statements. Retrieved June 9, 2014, from http://php.net/manual/en/pdo.prepared-statements.php

Regular-expressions.info. (n.d.). Regular Expressions Tutorial. Retrieved June 5, 2014, from http://www.regular-expressions.info/tutorial.html

Rutgers. (n.d.). Introduction to DNS. Retrieved July 24, 2014, from https://www.cs.rutgers.edu/~pxk/417/notes/content/ms_dns.pdf

SANS Institute. (2006, July). Packet Sniffing In a Switched Environment. Retrieved October 27, 2013, from https://www.sans.org/reading-room/whitepapers/networkdevs/packet-sniffing-switched-environment-244

West, M. (2011). Developing high quality data models. 1st ed. Burlington, MA: Morgan Kaufmann.

Windows Server. (n.d.). Understanding Reverse Lookup. Retrieved June 3, 2014, from http://technet.microsoft.com/en-us/library/cc730980.aspx

Wireshark. (n.d.). TCP Reassembly. Retrieved May 30, 2014, from http://wiki.wireshark.org/TCP_Reassembly

Yahalom, S. (2007, September). TCP Session Reconstruction Tool. Retrieved May 30, 2014, from http://www.codeproject.com/Articles/20501/TCP-Session-Reconstruction-Tool