Simulating Realistic Web Traffic using NS-3

Charles Carrington Scott
Hampton University
Mentor: Walid Younes, PhD, Network Simulation Group

Abstract

- The Network Simulation Group at Lawrence Livermore National Lab is working to build a realistic model of web traffic between a users machine and the web server. This will be performed by using a tool called NS-3. This tool is an open source software that allows you to create and simulate networking simulation on an extensible platform. This large platform is critical in order to fully simulate the many types of web traffic that exist. This module will increase the realism in ns-3 by making it easy to represent traffic between the user and servers, without actually representing actual content bytes.
- Understanding how applications will execute in a network context, characterizing the potential impact of these network threats, and predicting an effective defensive strategy are critical capabilities we hope to achieve with the help of ns-3.
- One important aspect to network simulation of this nature is understanding the many types of requests that are sent to and back from a web server. Many of these requests happen to be redirects to other web pages. Some of these redirects can be ads or even phishing attacks. Taking a look at these proportions between the two will help aid in the understanding of realistic web traffic.

Introduction

- To understand web traffic you must understand the types of requests that can exist in a topology (see Figure 1). When a user clicks on a website a HTML document is fetched along with hundreds of requests sent to and back from a web server to eventually populate what the user sees on that specific webpage. This HTML document acts as a shell to the website in order so that requests can complete the webpage. An issue that arises within these requests is that many redirects lie in a significant fraction of websites. Redirects bring on a whole new issue as far as simulation is concerned.
- These redirects that seem to lie in websites somewhat hide the true nature of what these websites consist of. Thus making it that much harder to reproduce real web traffic simulations within our modules. Taking a look at the proportions between requests to redirects in websites is the first step into tackling this tricky aspect of web simulation. Breaking down websites into certain categories (i.e. Technology, News/Media, Shopping etc.) will aid in the analysis of the data. Creating models of this collected data will give us a good breakdown of the amount of redirects that exists in today's web traffic requests.

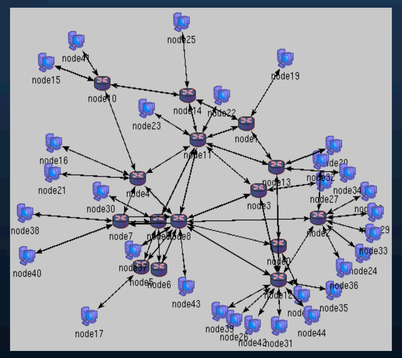

Figure 1: Topology Example

Methods

- The helpful resources that were used in order to collect the necessary information regarding each website include: alexa.com, httparchive.org and sitereview.bluecoat.com.
  - alexa.com provides extensive data about the most popular websites from a sample of millions of Internet users using over 25,000 different browser extensions. This tool was critical in terms of the most popular sites that should be analyzed.
  - httparchive.org collects and permanently stores the Web's digitized content. This source provides a complete breakdown of many if not most websites
  - sitereview.bluecoat.com categorizes each website according to the nature of the site (i.e. Technology, News/Media, and Shopping)

Figure 2: httparchive.org snapshot (Apple.com)

GET cnn_shdcamtt1l375.!	200	992 B	299ms
GET cnn_shdcafooter_20:	200	538 B	290ms
GET sprite_shades.gif	200	8.6 KB	289ms
GET Sprite_BT_master.gif	200	61 KB	300ms
GET ping.html?r=29388	200	3 B	821ms
GET 140701044811-pkg	200	6.2 KB	525ms
GET Type=count&ClientT	302	0	299ms
GET Type=count&ClientT	302	0	287ms
GET 140718161947-mh1	200	11.3 KB	279ms
GET 140718151635-hln-	200	4.2 KB	280ms
GET 140718132524-hln-	200	3.1 KB	281ms

Figure 3: httparchive.org snapshot (CNN.com)

- These tools were used in the collecting and the analyzing of this data. Next step, is parsing through the thousands of websites using the recourses above from conveniently accessible downloadable databases (i.e. httparchive.org). This was done by creating a simple Python script to parse through the data while extracting out the number of redirects and request for each website.
- The data is then plotted to aid in the understanding between the direct relationship between categories of websites as referred from sitereview.bluecoat.com

Results

The diagrams presented in this section were created based on the data extracted from downloadable databases according to the Python script. The data is divided between a number of different categories of websites according to sitereview.bluecoat.com.

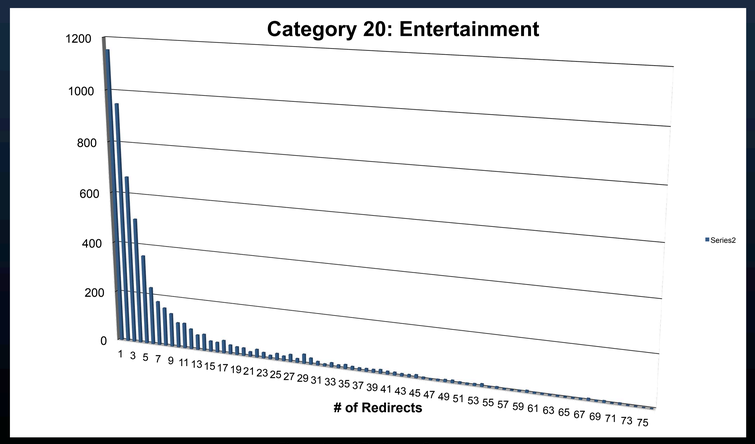

Figure 4: # of Redirects Occurrences

Results (Cont.)

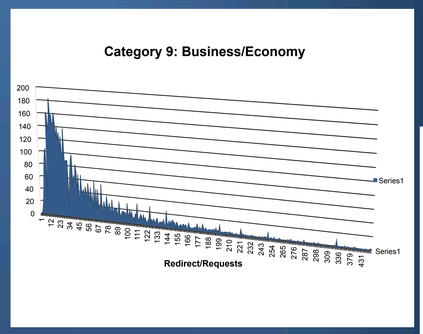

Figure 5: Redirects/Requests probability distribution

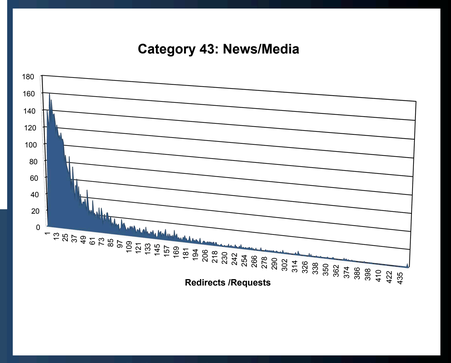

Figure 6: Redirects/Requests probability distribution

- The histogram in Figure 4 shows an example of Category 20's Redirects. During research it is found that all categories tested resembles this model. It can be concluded there is a high percentage of very low redirect counts in each of the tested categories.
- Figure's 5 and 6 shows the probability distribution between Redirects to Requests. This distribution was found by dividing each websites Redirects to Requests. Using this model above we can forecast the amount of Redirects each category may posses to Requests.

Discussion/Future Work

This research is far from over but the data presented shows that some things can be said in terms of requests and redirects within we traffic. Below is a summary of the findings:

- Most websites do not contain redirects
- At least 10% of the websites within each category will have redirects
- The produced models can provide an idea as to how many redirects will be performed for a given amount of requests
  - We can determine the average number of requests

Future work will consist of taking a more in depth look into the relationships between website categories and the proportions of redirects to requests. Also, taking a deeper look as to what these redirects consist of (i.e. malicious sites, website ads). Essentially a Monte Carlo simulation will be performed (on Figure's 4 & 5) which will aid in producing a simulated model that will closely resemble the realistic models produced here.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

LLNL-POST-658370