CS 590B Midterm Report

JACOB DOWNS, ELITZA NEYTCHEVA, and JOSHUA PIKOVSKY

ACM Reference format:
Jacob Downs, Elitza Neytcheva, and Joshua Pikovsky. 2016. CS 590B Midterm Report. 1, 1, Article 1 (January 2016), **??** pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Our project is focused on replicating the results of the FOCI 2014 paper titled Towards a Comprehensive Picture of the Great Firewallfis DNS Censorship. This paper presented an extensive investigation into the structure and logic of the Great Firewall of China, the primary technology in place to censor Internet traffic within the borders of China. The authors used external and internal access points, and provided examples that challenged the prior notion that censorship was limited only to internal Chinese traffic. The paper revealed that rather than simply censoring domestic traffic, the Great Firewall (or GFW) injected at the edges of Chinafis network. The paper thoroughly investigated what content the GFW blocked, a process aided by the fact that the GFW injected replies even in cases where the requested domain was invalid. In addition, their methodology involved sending requests with TTLs insufficient to reach their destination, which were nevertheless answered with a bogus response.

The authors determined that the GFW blocked domains based off criteria spanning from broad-scoped keywords, to domain names ending in exact matches. In total, 53,000 domains were found to be blocked from the 130 million sites taken from the Alexa lists. The paper also provided insight into the functionality of one GFW censorship node by using side-channels and other such methods to analyze and infer the behaviours of individual responses.

As described in our proposal, our plan for replication depended on several factors. If we were able to find the MaxMind geographic data, our project would involve querying DNS servers to determine the set of GFW resolvers, and testing for the size of the blocked domains list. The goal would be to investigate an individual node. If the MaxMind data was unavailable, then we would try to test for the size of the blocked domains list.

We were unable to obtain the MaxMind data, as it requires a paid membership, but we do have access to the Alexa top 1M sites

2 OUR PROGRESS

2.1 Midterm Progress

The first half of our plan for replication first obtaining the list of member domains in the 4 most populous top level domains along with the Alexa Top 1M domains. We were unable to obtain the MaxMind data set so we excluded it from our proceedings. That being said, we did successfully obtain an Alexa Top 1M domains list and zone files.

In order to facilitate the processing of data into more streamlined fashion, we built a parser which would take in our raw data sets and convert them into a standardized JSON form. This

2016. XXXX-XX/2016/1-ART1 $15.00
DOI: 10.1145/nnnnnnn.nnnnnnn

standardization will allow us to simplify our script for sending DNS requests and detecting blocked domain names.

We yet to create this script as of the time of the writing of this midterm report, however, all of the necessary information has been prepared for its creation. In order to eliminate the dependency issues in needing the parser script written before the packet sending script and the packet sending script before any of the data visualizations can be made, we agreed on standardized data formats which we could we would use to pass data between scripts before hand.

This format will be sure to map useful information to each of the tested domains such as whether his domain was blocked or not, what the TTL and IPID values are, and the response received. This will allow for correlations between whether a domain is blocked or not and the changes that occur in those selected values, allowing for characterization and analysis of the techniques employed by the GFW.

2.2 Remaining Work

Between now and the end of semester due date we will be writing the scripts to send the DNS request and generating visualizations. The visualizations we decide to use to present our data are likely to change depending on our results. For example, if we were to have results which indicate that the IPID is not an interesting field but the TTL seems to show some kind of strongly correlated change linked to whether or not a specific domain is blocked, we will favor creating visualizations for the TTL rather than the IPID.