Justin Ong

Professor Ujcich

7 December 2021

COSC 435 – Intro to Network Security

MP5

In a fictional world in which the only websites that exist are the five websites I have collected pcaps for (netflix.com, georgetown.edu, amazon.com, github.com, reddit.com), it is possible to differentiate which website a user is visiting even if they are using Tor. As the attacker, I could intercept the traffic between the user and the Tor network and capture the user's web traffic data. I can then analyze the traffic and compare it with my data collection results to determine which of the five websites they are most likely visiting.

When a user browses and arrives at the webpage, there are different files that must be transferred to the user before the webpage completely loads on the browser, such as images, videos, texts, and different scripts. This means that each website has different contents and by analyzing and defining the unique packet flow patterns, the contents of the website can be inferred. By looking at the mean, median, and standard deviations of the inter-packet arrival times, we can assess the delays in the packet flows for the website the user is visiting. More specifically, we can look at the inter-packet arrival times in both directions for when the source IP is the guard IP and for when the guard IP is the destination address. Additionally, for each direction, we can look at the length of packets, the total number of transmitted bytes, the directional changes in packet order, the unique packet lengths, and percentage of incoming bytes. After capturing and analyzing the user's data, we can simply compare it with the data we have already collected. We can compare the data numerically and graphically by creating histograms and charts, as well as logically thinking about the characteristics that each webpage has. For example, reddit.com has the most bytes in our collection which would make logical sense as the webpage opens up to a large amount of different threads that users like to see and read on the front page. And with Georgetown.edu, the front page loads with a large video playing and many large images as well. Lastly, because there are only five websites, the success rate of our comparison and differentiation would be higher than if the websites were not known.

Website	Mean (Guard -> Exit)	Mean (Exit -> Guard)	Median (Guard -> Exit)	Median (Exit -> Guard)	Standard Deviations (Guard -> Exit)	Standard Deviations (Exit -> Guard)	Total Number of Packets (Guard -> Exit)	Total Number of Packets (Exit -> Guard)	Total Number of Bytes (Guard -> Exit)	Total Number of Bytes (Exit -> Guard)
Amazon	0.018447732	0.0099698	0.001100063	0.000100136	0.039949245	0.10937385	688	508	824470	196492
Github	0.010330382	0.010757701	0.000779867	0.000139952	0.033536432	0.08266776	905	569	1228580	146188
Georgetown	0.00310309	0.000344513	0.000720024	0.000120163	0.009177054	0.00829153	5250	3131	7632068	438706
Netflix	0.006314161	0.004134198	0.000730038	0.000149965	0.019614897	0.11438009	1794	1086	2519836	220644
Reddit	0.003053844	0.001336458	0.000659943	0.000139952	0.010251944	0.035839543	9812	5808	14184378	1098596