A Large-scale Analysis of Content Modification by Open HTTP Proxies

Open HTTP proxies allows users within a network to store and forward internet services such as webpages usually in order to stay anonymous for restriction due to geological, censorship or content blocking. Routing traffic through a third party, always rises security issues whether the traffic been manipulated or code injection.

The traffic reaches to a proxy server can be interfered in order to make unauthorized alternations or to investigate furtively in an attempt to find information related to the user. Also monetizing web traffic of user by altering the content to inject ads, prompt user to download unwanted content or malwares and phishing attacks are another examples of what rogue proxy servers are capable of. More violent manipulation can also be applied such as malicious javascript code for mounting XSS, CSRF or DDoS attacks, injection of exploits against client side software or other actions which can result in user system to be endangered.

This paper analyzed a large scale amount of open HTTP proxies in order to determine the manipulation of user traffic or any other violation of privacy related issues as well as its extension. The proposed framework can also detect certain modification at the user traffic such as ad injection, redirecting to malware servers, user fingerprinting and user profiling. The results of this survey show that 5.15 percent of the targeted proxies, were employed in order for somehow modification of user content.

The content modifications that were considered in this work are those that enable the proxy provider to have monetary gain, affect user privacy and can result in system compromise. Also injection of ads and JavaScript codes for cryptocurrency mining, replacement of existing ads, attempt to steal user information and also SSL stripping techniques are considered.

In order to collect set of proxies, at the first step the keyword “HTTP proxy list” was used to make query through google search engine and after filtering irrelevant websites, 15 different websites that presents list of popular proxy were collected. Out of these collected websites, they manage

to crawl automatically 10 of them and 5 website manually for a period of two months. The manual and automated efforts result in 65871 unique proxies.

The complexity of differentiating between legal content modifications due to dynamic nature of content fir a specific websites and any other modification came by proxy services, make it necessary not to visiting real website rather using honeysites with different level of complexity. For testing purpose, the honeysites have been visited first through the proxies and second without them and then a comparison between the resulted contents have been made. It is also worth noticing that if the content of websites were modified in the expected way, no modification can be detected using this system. Also two level content modification clustering approach has been used for identifying injection that follows same pattern and grouping them together.

Among total number of proxies that was mentioned above, using TCP probes it was clear that 49444 proxies among them were alive for the period of this research. Also only 33.38 percentage of them which equals to 19473 proxies succeed in fetching and answering to requests that were made to test. By applying the comparison method, it was obtained that 7441 proxies modified the content of test webpages in some way. However it does not mean all of them are malicious and it was found in this research that 1004 proxies performed unwanted modification. It was also shown that public proxies that were obtained from free websites are more suspicious than those who were found from payed websites. It was interesting also that almost half of the proxies respond only less than 9 percent of the probes. Also only 28.46 percent of them have not being found in DNS based blacklist.

The behavior of rogue proxies were also categorize as follow: advertisements, fingerprinting, tracking and privacy leakage. It was also confirmed that some proxies change their behavior respecting the content that was requested by user, also ten percent of them modify the advertising IDs. In point of statistics, only 5.15 percent of proxies were found to perform some sort of malicious content modification.