

Understanding hidden websites deployed on Tor

Juha Nurmi

Tampere University of Technology

Email: juha.nurmi@ahmia.fi

Abstract—Tor is a software for anonymous TCP connections. This means that Tor enables anonymity to various Internet software. For instance, web servers can hide their location and web browsers can connect to these authenticated hidden services while the publisher and the viewer both stay anonymous. The publisher cannot be tracked down and the content cannot be censored. However, the research and measurements of hidden services is still thin. The aim of this paper is to introduce how we utilize our search engine implementation to understand hidden websites.

I. INTRODUCTION

Using the Tor network, it is possible to run web servers anonymously and without fear of censorship [1]. Servers configured to receive inbound connections through Tor are called hidden services (HSs): rather than revealing the real IP address of the server, a hidden service (HS) is accessed through the Tor network by mean of a virtual top level domain .onion [1].

In particular, we are interested in websites that operate as a hidden service. In this paper we call them hidden websites.

The published content is diverse [2]. Undoubtedly, some hidden websites are sharing pictures of child abuse, or operate as marketplaces for illegal drugs, including the widely known black market Silk Road. These few services are obviously controversial and often pointed out by critics of Tor and anonymity. On the other hand, vast number of hidden websites are devoted to human rights, freedom of speech, and information prohibited by oppressive governments.

Web search engines support finding web content. Because there were no search engines to search web content published using the Tor network, we built a working search engine for indexing, searching and cataloging content published inside the Tor network. Furthermore, we created an environment to share meaningful statistics, insights and news about the Tor network itself.

Ahmia provides the search and the access to hidden websites and believes that this is very important to the entire Tor network because we are efficiently enabling the diffusion and use of anonymous resources.

The whole search engine, Ahmia, is a free software and the source code is available online. This makes the research and our methods very transparent: Everyone is welcome to study our implementation.

In this paper we demonstrate how we can understand hidden service usage and how this reflects to our search engine design.

II. RELATED WORK

A. Tor2web

Tor2web is an HTTP proxy for Tor HSs designed initially by Aaron Swartz [3]. It aims at creating a network of proxies able to allow access to Tor HSs from the public internet. The software allows Tor hidden services to be reachable by means of a common browser and without the use of Tor client. Basically, it acts like a transparent proxy, translating the onion address into an HTTPS web URL [4].

In order to support tor2web.org, we maintain the Tor2web.fi proxy that enables people to connect to the .onion TLD with a regular web browser by replacing .onion part with .tor2web.fi. For example, <http://msydqstlz2kzerdg.onion> can be accessed using <http://msydqstlz2kzerdg.tor2web.fi>.

B. Access points to hidden websites

There are several hidden website directories similar to the early development of the web when Tim Berners-Lee maintained a list of websites. These sites, such as multiple hidden wikis, are specialized in linking to other hidden websites and categorizing those links. In addition, there are few other search engines crawling hidden websites.

C. Content and popularity analysis

The popularity and content of hidden services has already been measured and researched [2]. To do this, hidden services addresses has been collected by running Tor relays, the services has been port scanned, and in the case of HTTP services the content have been analyzed. Furthermore, the popularity of hidden services has been measured by looking at the request rate for hidden service descriptors by clients. However, this method exploits Tor’s design flaws and may cause privacy issues for hidden services. As such, it is not a sustainable way to measure the popularity and find hidden services.

The metrics measures the performance of the Tor network based on different data [5]. Some of these data sources are not sensitive at all, like properties and capabilities of a relay. Others are more sensitive, like statistics on fetched directory listings by country. As a result, Tor Metrics measures and shares the data about number of clients, bridges, relays, bandwidth, performance and diversity. While having lots of statistics on the Tor network, Tor metrics is not offering any insight to hidden services and their content.

First, the start point is to crawl the web content from the hidden services. Before that can be performed, a seed list of .onion domains has to exist. However, Tor technology does not offer a list of existing HSs. Therefore, the first seed list was originally gathered from the sites that were listing .onion URLs.

Another problem with current linking structure is that typical search ranking algorithms are based on linking between websites. Because the linking between hidden websites is thin, normal ranking algorithms perform poorly.

Similarly, using the search index of hidden websites, we are locating malicious software sites to inform security firms, child pornography sites to filter them out, and immediately after the international law enforcement operation, Operation Onymous, the list of sites seized by them.

Moreover, very informative measurement is the total number of public WWW backlinks pointing to known hidden websites. We fetch this information automatically from the popular normal search engines.

REFERENCES

-

[illegible]

By Tor2web average visits By public WWW backlinks

Fig. 3. The most popular websites according to average Tor2web proxy visits per day and the most popular websites according to number of backlinks from the public WWW.