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.

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.

mds
December 05, 2014

II. Related work

A. Tor2web

Tor2web is an HTTP proxy for Tor HSs designed initially by Aaron Swartz[4]. 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[5].

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 is 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[3]. 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.

III. Finding, ranking and understanding content

First, the start point is to crawl the web content form the hidden services. Before that can be performed a seed list of .onion domains is used. 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.

Unfortunately, this method finds only those new .onion sites which are linked to those .onion pages which are already indexed. Moreover, only few hidden websites links to other hidden websites and there is no linking to every .onion. As a result, we cannot find all hidden sites. We visualized this problem by generating a SVG image of the crawling paths (see figure 1). More visualizations material is available on https://ahmia.fi/static/visuals/.

Another problem 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.

We would like to show a glance of the hidden website content in general. Ahmia produced a word cloud visualization of the front pages of hidden websites (see figure 2).

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.

With the Tor2web software developers we introduced a HS discovery function to Tor2web software. This means that Tor2web is gathering a list of the visited .onion websites and the visit counts and search engines can download this list. In this way we can find new .onion domains and use their popularity information. This means that we can find new hidden websites and measure their popularity.

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.

These popularity metrics are illustrated in the figure 3.

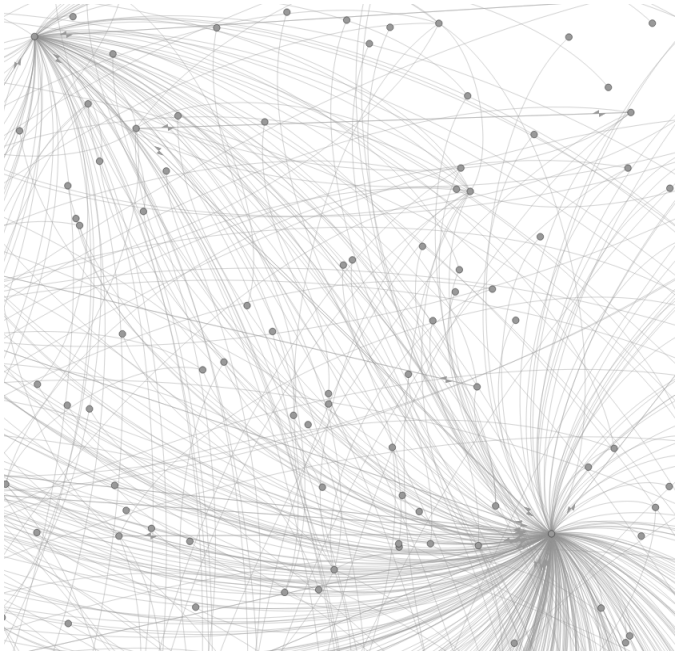

Fig. 1. A part of the visualization of the linking structure of hidden websites. Few sites gather lists of other .onion domains while the most of the sites have no linking to other .onion domains.

Fig. 2. A word cloud that represents the most popular text content. These are the most common words on the front pages of hidden websites.

By Tor2web average visits	By public WWW backlinks
216003 pinkmethuylnenlz.onion	24538 strngbxhwyuu37a3.onion
22873 t54cjs4qc2r4bn63.onion	3252 kpvz7ki2v5agwt35.onion
13223 3qwajq5p5pfsi3sw.onion	2852 silkroad6ownowfk.onion
13064 h3vf5leilsvjiqlx.onion	1830 silkroad5v7dywlc.onion
12799 torbookdjwhjnju4.onion	1520 3g2upl4pq6kufc4m.onion
12239 svcz25e3m4mwlauz.onion	1510 am4wuhz3zifexz5u.onion
11414 npdaaf3s3f2xrmlo.onion	1410 grams7enufi7jmdl.onion
7756 64ansq6xm5mmsb3a.onion	1350 xmh57jrzrnw6insl.onion
7266 juvatztgkapzrp2o.onion	1034 silkroadvb5piz3r.onion
6530 girlshtjireiazwm.onion	1020 agorahooawayyfoe.onion

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.

References

[1] Dingledine, Roger, Nick Mathewson, and Paul Syverson. Tor: The second-generation onion router. Naval Research Lab Washington DC, 2004.

[2] Biryukov, Alex, Ivan Pustogarov, and Ralf-Philipp Weinmann. "Content and popularity analysis of Tor hidden services." arXiv preprint arXiv:1308.6768 (2013).

[3] Loesing, Karsten, Steven J. Murdoch, and Roger Dingledine. "A case study on measuring statistical data in the tor anonymity network." Financial Cryptography and Data Security. Springer Berlin Heidelberg, 2010. 203-215.

[4] In Defense of Anonymity. Aaron Swartz. http://www.aaronsw.com/weblog/tor2web

[5] Tor2web software project. HERMES Center for Transparency and Digital Human Rights. http://logioshermes.org/home/projects-technologies/tor2web/