Understanding hidden websites deployed on Tor

Juha Nurmi
Tampere University of Technology
The founder and a developer of Ahmia – Tor hidden service search

juha.nurmi@ahmia.fi

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, finding web content is laborious without an efficient search engine and therefore a search engine is needed for the Tor network.

The aim of this paper is to introduce how to use our search engine implementation to understand hidden website.

Anonymity is an important right in order to support freedom of speech and defend human rights. An Internet user can use range of tools to hide ones identity[1]. Among these, the most popular tool is Tor. It has a large number various users, including ordinary citizens concerned about their privacy, corporations who do not want to reveal information to their competitors, and law enforcement and government intelligence agencies who need perform operations on the Internet without being noticed[2]. Further, human rights activist and journalist are communicating anonymously using Tor to protect their lives[3].

The Tor network is considered to be well studied and very secure communication network[2]. According to top secret NSA documents disclosed by a whistleblower Edward Snowden[4], who is a former Central Intelligence Agency (CIA) employee, and former National Security Agency (NSA) contractor, the Tor network protects its users against surveillance. The NSA even wrote in their top secret documents that **Tor is "the King of high secure, low latency Internet Anonymity"**.

In addition, using the Tor network, it is possible to run web servers anonymously and without fear of censorship[5]. 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**[5].

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

As a result, the published content is diverse[6]. 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.

However, it is laborious to find content, published using hidden websites, without a search engine. The Tor technology is designed to offer a method to register .onion domains and to obtain anonymous TCP connections while it lacks a method to search the actual content. Obviously this is a problem of the applications layer and Tor is operating on the transport layer. Furthermore, even though the WWW is designed for information sharing it lacks a built-in mechanism to search the published information.

Web search engines enable finding the 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.

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 needed. 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 (figure 1). More visualizations material is available on https://ahmia.fi/static/visuals/.

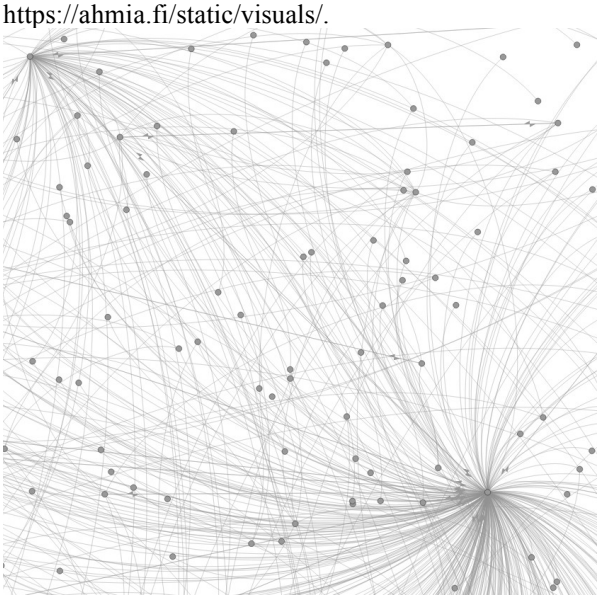

Figure 1. *A part of the visualization of the linking structure of hidden websites. Few sites gather lists of other .onion domains while the most of the sites have no linking to other .onion domains.*

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.

Figure 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

Figure 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.*

1. References

[1] Goldschlag, D., Reed, M., and Syverson, P., Onion routing, Communications of the ACM, Location, 1999.

[2] Dingledine, R., Mathewson, N., and Syverson, P., Deploying low-latency anonymity: Design challenges and social factors, Security & Privacy, IEEE, , 2007.

[3] , , , , .

[4] , Tor: 'The king of high-secure, low-latency anonymity', theguardian.com, , Friday 4 October 2013 15.49 BST.

[5] Dingledine, R., Mathewson, N., and Syverson, P., Tor: The second-generation onion router, Naval Research Lab, Washington DC, 2004.

[6] Biryukov, A., Pustogarov, I., and Weinmann, R. P., Content and popularity analysis of Tor hidden services, arXiv preprint, , 2013.