A Deep Dive into the Dark Web

Research project as part of the Security and Network Engineering master at UvA

Coen Schuijt: cschuijt@os3.nl

February 10th, 2019

Abstract

The size of the web is ever increasing. The web reaches far beyond the part that is indexed and searchable by search engines, also known as the surface web. This research focuses on measuring the surface web and a specific part of the deep web – the dark web – that is accessible by means of the TOR protocol. The amount of pages on the surface web were obtained via a literature study, while the mean page size was measured based on samples gathered from several search engines. An overlap analysis was conducted to gain insights in the total amount of pages accessible through TOR. A script was built in order to measure average page sizes without actually accessing those, due to ethical considerations. The results show that the surface web is roughly between 16 and 193 pebibytes, while TOR is between 0.3 and 0.8 pebibytes – which results in a TOR/surface-ratio of approximately 0.6%.

1 Introduction

The visible part of the web – also called the surface web – is the part of the web that is being crawled and indexed by search engines [1, 2]. The deep web, on the other hand, is characterized as the part of the web that has not been indexed by search engines. The dark web is part of the deep web and often only accessible via special software or authentication [3]. There exists a huge amount of web pages that cannot be accessed directly, but only via dynamically issued queries to search interfaces of databases [4]. Some people claim that the 'surface' web is about 4% of the internet, while the 'deep' and 'dark' web combined are about 96% of the internet [5, 6, 7].

These claims are often based on a research paper by Bergman from 2001 [8]. One of the objectives of that research was to quantify the size of the deep web [8]. The problem is that the amount of information available on the web is ever increasing, though only a small percentage is indexed. Because search indexes are used to find and deliver (correct) information, it is therefore necessary to gain insights into the percentage of the web that is indexed and covered by these search indexes [9, 10].

The focus of this research is to either verify or reject the statements being made about the size ratio of the surface web and the part of the deep web that is accessible via the TOR protocol over HTTP.

The remainder of this paper is structured as follows. The research questions are outlined in Section 1.1, Section 2 provides an overview of related research, Section 3 elaborates the methodology, Sections 4 and 5 contain the results and the conclusion, Section 6 consists of the discussion and Section 7 mentions future work.

1.1 Research question

The main research question for this project is defined as follows:

> **"What is the size ratio of the deep web that is accessible over the TOR protocol as compared to the surface web?"**

In order to answer this question, the following subquestions need to be answered:

1. What are the definitions for surface web, deep web and dark web?
2. How to estimate the total size of the web based on the size of a subset?
3. What metrics are applicable for measuring and defining the size of (a subset of) the web?

1.2 Parts of the web

To clarify the main research question, Figure 1 depicts the structure and various sections of the web: The web as a whole consists of the surface web (green), which indexed a part of the deep web (orange). The dark web (red) is part of the deep web.

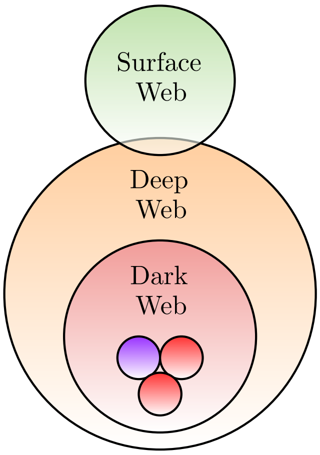

Figure 1: *Schematic overview of the web.*

The dark web, again, consists of different components, as shown in Figure 2 (Left), each accessible by means of special software or authentication [3]. Figure 2 (Right) shows the parts of the web that are compared in this research.

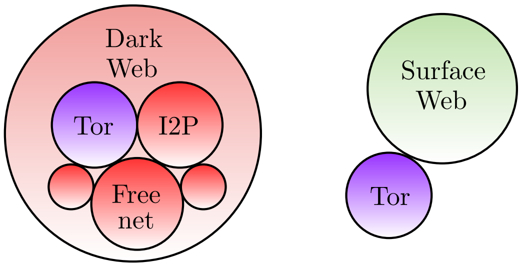

Figure 2: *Overview of various parts of the dark web* (Left) *and the comparison for this research* (Right).

2 Related research

Van den Bosch et al. provide an extensive overview of research regarding measuring the web [10], specifically the surface web. Most notable is the research by Bharat and Broder's from 1998 [11], who provided a method for estimating the size by randomly selecting pages from a search index and verifying whether they occur in another index and vice versa [10]:

- *Page selection* - a lexicon of about 400.000 words was created, from which a total of 35,000 random queries with a length of 6 to 8 words were derived. Four search engines were queried and a random page was selected from the first 100 pages.
- *Verifying* - the page from the previous step was queried against other indexes. This was done by "taking the top *k* most discriminant terms from each randomly selected page" – creating a so called *strong query* – [10] and verifying whether a result from that query matched the original URL from the sample.

Further effort has been made in order to improve on the aforementioned approach, mostly focusing on the sampling method. According to [10], using Bharat and Broder's method as a starting point "can be problematic because not every page has the same probability of being sampled" using this approach. Van den Bosch et al. use a different approach, by means of extrapolating search results based on a known corpus.

Bergman was among the first to estimate the size of the deep web [8]. Chen wrote a book covering a lot of information regarding data mining the dark side of the web and its uses [3]. Furthermore, various efforts were made to crawl parts of the deep and dark web. Gupta and Bhatia provided a comparative study of a lot of those crawlers, including their strengths and limitations [12].

3 Methodology

This section first describes the general approach, followed by the approach for measuring the size of the surface web and, lastly, the approach for measuring the size of the deep web through TOR.

3.1 General approach

As the main goal of this research is to compare the size of the surface web to the part of the deep web that is accessible over the TOR protocol, the size of both parts had to be measured. The size of a part of the web can be expressed in various metrics, such as the amount of URLs [11] or as the size when storing the content of the web pages on disks [8]. The main approach for comparing the aforementioned parts of the web consists of the following steps:

1. Determine the amount of pages on the surface web
2. Determine the average page size of websites on the surface web
3. Determine the amount of pages on the deep web that are accessible through TOR
4. Determine the average page size of websites on the deep web accessible through TOR
5. Calculate sizes and ratio

The calculation for the ratio r could then be depicted as:

$$\mathrm{r} = \frac{(\#\ \text{TOR pages} \times \text{mean TOR page size})}{(\#\ \text{surface pages} \times \text{mean surface page size})} \times 100\%$$

3.2 Measuring the surface web

The amount of pages that exist on the surface web were determined based on a literature study. In order to estimate the average page size of the surface web, three search engines (Bing, Google and Yahoo) were queried with the 27 *pivot words* as mentioned by [10] – representing various frequency ranks, from high to low – to gather representative page size results.

3.2.1 Size estimation (surface web)

Let $\mathrm{S_L}$(surface) be the lower bound and $\mathrm{S_U}$(surface) be the upper bound amount of websites on the surface web. For each of the 27 pivot word $\mathrm{w_i}$, 3 search engines $\mathrm{e_i}$ were queried, after which 10 random pages $\mathrm{p_i}$ were selected and saved.

Let A be the set with all samples $\mathrm{x_i} \in \{\mathrm{x_1}, \dots, \mathrm{x_n}\}$. The size of A, denoted as N, is the result of $\mathrm{w_i} \times \mathrm{e_i} \times \mathrm{p_i}$, which is equivalent to $27 \times 3 \times 10 = 810$. The mean page size $\overline{\mathrm{x}}(\mathrm{p})$ could then be calculated based on this collection of samples, resulting in the following equation:

$$\overline{\mathrm{x}}(\mathrm{p}) = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{x_i}$$

The sample standard deviation s was calculated as follows:

$$\mathrm{s} = \sqrt{\frac{1}{\mathrm{N}-1} \sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathrm{x_i} - \overline{\mathrm{x}})^2}$$

Based on this information, an approximation could be made for the lower bound page size $\mathrm{p_L}$(surface) and upper bound page size $\mathrm{p_U}$(surface) for pages on the surface web.

The lower bound total web size $\mathrm{T_L}$(surface) and upper bound total web size $\mathrm{T_U}$(surface) were then calculated by the products of every combination of the lower and upper bound amount of web pages [$\mathrm{S_L}$(surface) and $\mathrm{S_U}$(surface)] and the upper and lower bound page sizes [$\mathrm{p_L}$(surface) and $\mathrm{p_U}$(surface)].

3.2.2 Query Parameters

In order to make sure that random samples were selected, for each of the search engines a query parameter was used to start showing results with a random offset. For Bing, this was accomplished by adding the '&first=' query parameter, resulting in the following URL: `https://www.bing.com/search?q=WORD&first=RAND`. For searches with Google, the '&start=' query parameter was used, which resulted in the following URL: `https://www.google.com/search?q=WORD&start=RAND`. Lastly, for searches within Yahoo, the '&b=' query parameter was used, resulting in the following URL: `https://search.yahoo.com/search?p=WORD&b=RAND`.

Within each URL, 'WORD' depicts the pivot word w and 'RAND' depicts a random value between 0 and 500.
An effort was made to download the pages from the surface web in an automated way, by making use of the "Bing Search APIs v7" for Bing and the "Google Custom Search API" for Google. However, the Google Search API limits the maximum amount of queries per day. Neither was it possible to mimic the saving of a web page in order to download a full copy. To account for these limitations, all pages were saved by hand.

3.3 Measuring the deep web

The part of the deep web that is accessible over the TOR protocol was measured in a similar way as the surface web. However, no estimates for the amount of websites were available. In order to gain insights in the total amount of .onion sites, a scraper was built to gather as much unique links as possible.

3.3.1 Experimental Setup

In order to access the web pages via TOR, the test setup as depicted in Figure 3 was used. A physical machine (`Workstation`) was first connected to a VPN node (`VPN`). The physical machine hosted two virtual machines in order to access the TOR network securely. One of the guest machines (`Kali`) was connected with a TOR gateway (`Whonix`) via an internal network. The Whonix VM connected used a NAT interface of the Workstation. The TOR data itself consists of several layers of encryption.

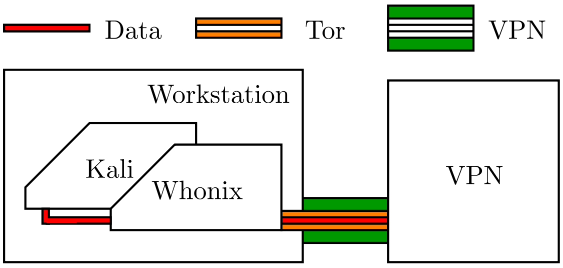

Figure 3: *Schematic overview of the experimental setup.*

3.3.2 Gathering procedure

Initially, several onion lists that are publicly accessible via the surface web were crawled for .onion links. These "seed URLs" were added to a file called '`surface-seeds.lst`'. Additionally, several deep web links with an .onion extension were added to a file called '`onion-seeds.lst`' to be used as seed URLs. These seed URLs were then crawled, after which all onion links were parsed and saved. The following scripts were used, with their functionalities [13]:

- `surface_scraper.py` and `onion_scraper.py` – first reads the seed URLs from the file `url-seed-file-[surface|onion].lst`, respectively, and parses the seed URLs. For each seed URL, the HTML file is downloaded and stored as a raw file. If a file already exists, the content of the (sub)page is appended to that file. Then, a regular expression is used to parse the .onion links in each of the raw files and writes those links to parsed files.
- `measure_overlap.py` – creates a list with unique domains in the folder with parsed files. Then, all files for the same domain with different dates are grouped together. A set with all unique .onion sites for each seed URL is created. All possible combinations for all unique sets are mapped: $\langle A, B\rangle, \ldots, \langle A, G\rangle, \langle A, B, C\rangle, \ldots, \langle E, F, G\rangle, \ldots, \langle A, B, C, D, E, F, G\rangle$. For all combinations the overlap is measured.
- `create_sample.py` – this script requires a list with samples, an output file name, as well as a count as inputs. It reads the contents of the specified list of samples and returns the specified amount (count) of random samples.
- `measure_mean_size.py` – first reads the list of samples provided. Then checks whether the URLs are responsive asynchronously. The page size, including all page content, is saved after which the total size is measured and returned.

3.3.3 Page amount estimation (TOR)

First, the seed URLs available via the surface web were crawled. For each seed URL, a unique list with onion addresses was created. Then, the overlap between each of the lists was measured. The same procedure was repeated for the seed URLs that were gathered via the TOR protocol. Lastly, a mixed overlap analysis was performed, measuring the overlap of the two largest lists from both the surface web and TOR.

Similar to the overlap analysis being used in [8, 11], based on the overlap ratio of two subsets, the total amount of web pages available via the TOR protocol could be estimated.

Let $S_L(\mathrm{tor})$ be the lower bound and $S_U(\mathrm{tor})$ be the upper bound amount of pages on the deep web that are accessible through TOR. Let F_{sub} and G_{sub} be two random subsets being used within the pair-wise overlap analysis as obtained via the gathering procedure mentioned earlier. In order to estimate the total size of S_{TOR}, one would first have to calculate the overlap, or the amount of links that exist in both sets, denoted $F_{sub} \cap G_{sub}$. Then, the estimate of the fraction of the total size as covered by F_{sub}, is calculated as $G_{sub}/(F_{sub} \cap G_{sub})$. The total size S_{TOR} could then be estimated based on dividing the total size of F by this ratio. A schematic overview is given in Figure 4 – where F depicts F_{sub} and G depicts G_{sub}.

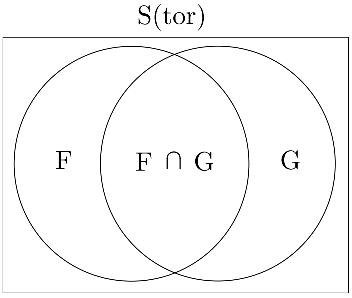

Figure 4: *Schematic overview of overlap analysis.*

3.3.4 Page size estimation (TOR)

After scraping the seed URLs, both from the surface web as well as the ones available through TOR, all onion links were aggregated into one single file. As opposed to saving page samples based on pivot words, the mean page size $\bar{y}(p)$ for pages accessible via TOR, were obtained from the total set of gathered data.

Let B be the set with all samples $y_i \in \{y_1, \ldots, y_m\}$.

The size of B, denoted as M, is obtained as a random subset of S_{TOR}. The mean page size $\overline{y}(p)$ for pages accessible via TOR could then be calculated based on this collection of samples, resulting in the following equation:

$$\overline{y}(p) = \frac{1}{M} \sum_{i=1}^{M} y_i$$

The sample standard deviation t was calculated as follows:

$$t = \sqrt{\frac{1}{M-1} \sum_{i=1}^{M} (y_i - \overline{y})^2}$$

3.3.5 Verification of results

To make sure that the estimations by the script would reflect the actual page sizes on the dark web, a series of tests were conducted. Due to the sensitive content on the dark web, using a similar approach for collecting samples as used with the surface web (saving pages by hand) is not possible. Therefore, the accuracy of the results gathered by the `measure_size.py` script were verified with sites known to not contain sensitive data. The script was developed, tested and optimized according to the following phases:

1. White Box – The script was developed based on a selected set of pages, which are known to not contain sensitive information. The page was saved by hand and the script was optimized so that the estimated page size was as close as possible to the actual page size.
2. Grey box – Another series of selected pages, known to not contain sensitive data, but not used in prior tests, was used to verify the script's estimating accuracy on unknown pages. The same procedure for saving and optimization as with item 1 was applied to optimize the script accuracy.
3. Black box – The script was tested on a random set of samples as generated with the `create_samples.py` script.

Based on this information, an approximation could be made for the lower bound page size p_L and upper bound page size p_U for pages on the surface web.

The lower bound total web size $T_L(tor)$ and upper bound total web size $T_U(tor)$ were again calculated by the products of the amount of web pages S_{TOR} and the upper and lower bound page sizes (p_L and p_U).

4 Results

Section 4.1 contains the amount of pages, the mean page size and the estimations of the surface web. Section 4.2 contains the amount of pages, the mean page size and the estimations for TOR. Section 4.3 describes the ratio of both parts of the web.

4.1 Surface web

Research by van den Bosch et al. [10] has estimated the size of the web to be at least 6 billion pages as of Thursday, January 24th, 2019, which will be used as a lower bound approximation $S_L(surface)$ of the total amount of web sites. The upper bound estimations by the same researcher [14] go up to approximately 53 billion pages as of Thursday, January 24th. This will be used as the upper bound estimation $S_U(surface)$ of the total amount of websites.

4.1.1 Average Page Size

The average page sizes as calculated by retrieving web pages via various search engines are as follows:

- Bing results appeared to be ≈ 795 MiB for 270 sites, resulting in a mean page size of ≈ 3,016 KiB.
- Google results appeared to be ≈ 982 MiB for 270 sites, resulting in a mean page size of ≈ 3,723 KiB.
- The Yahoo results ended up being ≈ 941 MiB for 270 sites, resulting in a mean page size of ≈ 3,569 KiB.

The 810 samples had a total size of ≈ 2755 MiB, with a mean value of ≈ 3483 KiB and a margin of error of ≈ 529 KiB using a 95% confidence interval. This results in a lower bound page size $p_L(surface)$ of ≈ 2955 KiB and a upper bound page size $p_U(surface)$ of ≈ 4012 KiB.

4.1.2 Total size of surface web

The lower bound total web size $T_L(surface)$ and upper bound total web size $T_U(surface)$ were calculated by the products of all combinations of the lower and upper bound amount of web pages [$S_L(surface)$ and $S_U(surface)$] and the upper and lower bound page sizes [$p_L(surface)$ and $p_U(surface)$], as noted in Table 1:

- The lower bound total web size $T_L(surface)$ is between 16.12 and 21.89 pebibytes.
- The upper bound total web size $T_U(surface)$ is between 142.43 and 193.40 pebibytes.

4.2 Deep Web

First, a total of 7 onion lists – that are accessible via the surface web – were crawled for .onion links. This resulted in an initial data set of 24,446 unique links, gathered in the period between January 15th and 20th 2019. During the period from January 21st until January 23th, another 41,694 links were gathered. This resulted in a total set of 46,779 unique links.

4.2.1 Overlap analysis

Initially, the pair-wise overlap of the 7 sources from the surface were measured. The truncated addresses and their respective source are mentioned in Table 2, resulting in the

Web Size	Page Size	Equation	Result (bytes)	Result (pebibytes)
$S_L(surface)$	$\overline{x}(p_L)$	$6 \times 10^9 \times \approx 2955$ KiB	$\approx 1.82 \times 10^{16}$ bytes	≈ 16.12 PiB
$S_L(surface)$	$\overline{x}(p_U)$	$6 \times 10^9 \times \approx 4012$ KiB	$\approx 2.47 \times 10^{16}$ bytes	≈ 21.89 PiB
$S_U(surface)$	$\overline{x}(p_L)$	$53 \times 10^9 \times \approx 2955$ KiB	$\approx 1.60 \times 10^{17}$ bytes	≈ 142.43 PiB
$S_U(surface)$	$\overline{x}(p_U)$	$53 \times 10^9 \times \approx 4012$ KiB	$\approx 2.18 \times 10^{17}$ bytes	≈ 193.40 PiB

Table 1: *Estimations of total surface web size, based on lower bound and upper bound web page sizes, as well as upper and lower bound of the total amount of web sites.* (*The values depicted are rounded to the second decimal.*)

Venn diagram as depicted in Figure 5. The full addresses can be found in Appendix A.

ID	Seed	Web	Sites
1	thehiddenwiki.org	Surface	173
2	ahmia.fi	Surface	20798
3	github.com/alecmuffett	Surface	127
4	onions.danwin1210.me	Surface	5352
5	github.com/agentWotes	Surface	73
6	guthub.com/kenorb	Surface	3202
7	deep-weblinks	Surface	350

Table 2: *Seed lists found on the surface web, including their respective sizes.*

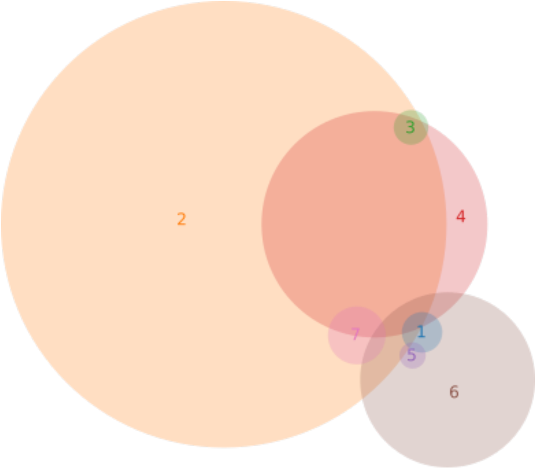

Figure 5: *Overlap analysis of onion lists from the surface web, gathered during the period from January 15th until January 20th, 2019. The labels represent the sites as mentioned in Table 2.*

Then, the overlap of the 7 additional sources found through TOR web were measured. The truncated addresses and their respective size are listed in Table 3, resulting in the Venn diagram as depicted in Figure 6. The full addresses can be found in Appendix A.

ID	Seed	Web	Sites
A	visitorfi5kl7q7i.onion	Tor	1135
B	underdj5ziov3ic7.onion	Tor	41459
C	jh32yv5zgayyyts3.onion	Tor	255
D	wikitjerrta4qgz4.onion	Tor	287
E	torlinkbgs6aabns.onion	Tor	167
F	donio(...)q2vead.onion	Tor	4461
G	torvps7kzis5ujfz.onion	Tor	879

Table 3: *Seed lists accesses through TOR, including their respective sizes.*

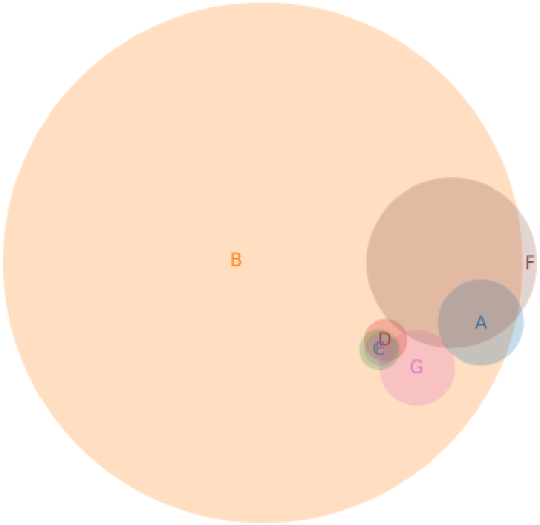

Figure 6: *Overlap analysis of onion lists from TOR, gathered during the period from January 21st until 23rd, 2019. The labels represent the sites as mentioned in 3.*

Table 4 shows the sources being used for the mixed overlap. For each of these seed lists, the amount of overlap with the other lists is measured as part of a pair-wise overlap analysis. The ID's are kept consistent with the ID's being used in Table 2 and Table 3. The resulting Venn diagram is depicted in Figure 7.

ID	Seed	Web	Sites
2	ahmia.fi	Surface	20798
B	underdj5ziov3ic7.onion	Tor	41459
4	onions.danwin1210.me	Surface	5352
F	donio(...)q2vead.onion	Tor	4461

Table 4: *Various seed lists, including the part of the web the source was found and their respective sizes.*

The results of the mixed overlap are included in Table 5. For every pair **A** and **B** the the amount of elements that exist in both sets is measured, denoted as $A \cap B$. Then, for every entry, the overlap is divided by the total size of set B, resulting in the fraction of the total size. For every combination, set A is divided by the ratio in order to estimate the total size of S_{TOR}. The ratio varies from ≈ 0.41 up to ≈ 0.99.

The site `onions.danwin1210.me` listed 4400 onion addresses, from which 227 were online more than 7 days (5.2%), as measured on January 28th, 2019, . On February 3rd, 214/4387 sites were online longer than 7 days (4.9%). The site `haystakvxad7wbk5.onion` claims to have indexed 1.5 billion onion links.

A	A count	B	B count	$A \cap B$	Ratio	Estimation
2	20798	B	41459	17108	$17108/41459 = 0.41$	$20798/0.41 = 50401$
2	20798	4	5352	4511	$4511/5352 = 0.84$	$20798/0.84 = 24675$
2	20798	F	4461	3700	$3700/4461 = 0.83$	$20798/0.83 = 25075$
B	41459	4	5352	5143	$5143/5352 = 0.96$	$41459/0.96 = 43143$
B	41459	F	4461	4250	$4250/4461 = 0.95$	$41459/0.95 = 43517$
4	4461	F	4461	4423	$4423/4461 = 0.99$	$4461/0.99 = 4499$

Table 5: *Estimations of total TOR web size, based on a pair-wise overlap analysis of the two largest subsets of the surface web and TOR. (The values depicted are rounded to the second decimal.)*

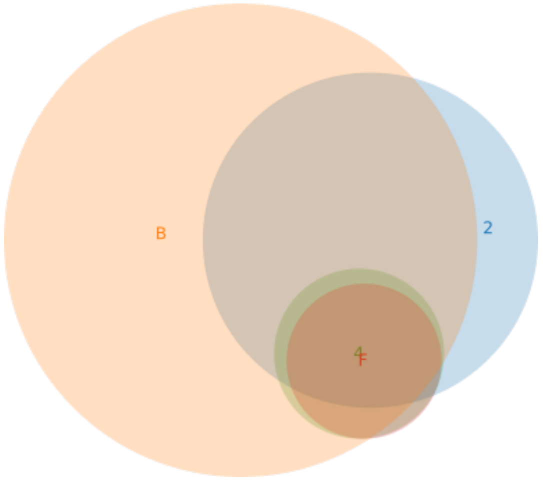

Figure 7: *Mixed overlap analysis of onion lists. The labels correspond to Table 4.*

4.2.2 Average Page Size

In order to measure the average page size of pages available over the TOR protocol, a total of 99 pages were measured. The `measure_size.py` script was created and optimized in three phases, as described in Section 3.3.5. Table 6 shows an overview of the links being measured, the phase they were measured in, the size as approximated by the script (in bytes), the actual size after manual inspection (in bytes) and the offset between approximation and the actual result.

After manually inspecting the results for the site `xmh57jrzrnw6insl.onion`, it turned out that this page was loading another page containing 19 gif images. When measuring the page (`xmh57jrzrnw6insl.onion`) and add the HTML of the loaded page, this resulted in a page size of 11113 bytes, which means the approximation was only 101.5% off the actual size.

The 99 samples had a total size of ≈ 22 MiB, with a mean value of ≈ 227 KiB and a margin of error of ± ≈ 26 KiB using a 95% confidence interval. This results in a lower bound page size $\mathrm{p_L(tor)}$ of ≈ 200 KiB and a upper bound page size $\mathrm{p_U(tor)}$ of ≈ 253 KiB.

4.2.3 Total size of TOR

Taking the amount of 1.5 billion pages as mentioned by the haystak website as a starting point, the total amount of websites can be extrapolated based on the ratios as mentioned in 5. Multiplying the amount of websites with the maximum ratio results in the lower bound approximation of amount of websites $\mathrm{S_L(tor)}$, while multiplying with the minimal ratio results in the upper bound approximation $\mathrm{S_U(tor)}$ of amount of websites:

- $\mathrm{S_L(tor)}$ is $(1.5 \times 10^9)/0.99 \approx 1.5 \times 10^9$ sites
- $\mathrm{S_U(tor)}$ is $(1.5 \times 10^9)/0.41 \approx 3.6 \times 10^9$ sites

The lower bound total web size $\mathrm{T_L(tor)}$ and upper bound total web size $\mathrm{T_U(tor)}$ were calculated by multiplying the lower and upper bound amount of web pages available through tor [$\mathrm{S_L(tor)}$ and $\mathrm{S_L(tor)}$] with the lower and upper bound approximated web page sizes [$\mathrm{p_U(tor)}$ and $\mathrm{p_L(tor)}$], similar to the calculations for the surface web:

- $\mathrm{S_L(tor)} \times \mathrm{p_L(tor)} \approx 3.10 \times 10^{14} \approx 0.28$ PiB
- $\mathrm{S_L(tor)} \times \mathrm{p_U(tor)} \approx 3.92 \times 10^{14} \approx 0.35$ PiB
- $\mathrm{S_U(tor)} \times \mathrm{p_L(tor)} \approx 7.46 \times 10^{14} \approx 0.66$ PiB
- $\mathrm{S_U(tor)} \times \mathrm{p_U(tor)} \approx 9.42 \times 10^{14} \approx 0.84$ PiB

This results in a lower bound total web size $\mathrm{T_L(tor)}$ between 0.28 and 0.35 PiB. The upper bound total web size $\mathrm{T_U(tor)}$ is between 0.66 and 0.84 PiB.

4.3 Ratio

When taking the mean results for the size estimations of the surface web, the mean total size T(surface) is 93.46 PiB. Doing the same for the estimations of TOR, the mean total size T(tor) is 0.53 PiB. This would mean that the size freely accessible part of the dark web through TOR over HTTP is about 0.56% of the size of the surface web.

5 Conclusion

The surface web appears to be at least between 6 and 53 billion pages as measured on Thursday, January 24th, 2019. The average page size on the surface web was measured to be between 2955 and 4012 KiB, as obtained by 810 random samples. The part of the dark web that is accessible over HTTP(S) through TOR is approximated to be between 1.5 and 3.6 billion pages. The average page size is lower, measure to be between between 200 and 253 KiB, as obtained by 99 random samples. The final comparison resulted in a TOR/surface ratio of approximately 0.53/93.46 PiB, which is about ≈ 0.6%.

Site	Phase	Script	Manual	Offset
torlinbgs6aabns.onion	White	44504	44105	100.9% (+0.9%)
jh32yv5zgayyyts3.onion	White	74306	72681	102.2% (+2.2%)
onions.danwin1210.me	White	51834	58177	89.1% (-10.9%)
onionlstmjc7qkmj.onion	White	144063	144132	100.0% (0.0%)
xmh57jrzrnw6insl.onion	Grey	11448	7256992	0.2% (-99.8%)
5plvrsgydwy2sgce.onion	Grey	303417	205195	147.9% (+47.9%)
uj3wazyk5u4hnvtk.onion	Grey	21589	22864	94.4% (-5.6%)
haystakvxad7wbk5.onion	Grey	428333	428273	100.0% (0.0%)
qhhunyjzmdyx4i4d.onion	Grey	11402	10796	105.6% (5.6%)
libraryqtlpitkix.onion	Black	179	283	63.3% (-36.7%)
wolfmu4yjw3srihs.onion	Black	434874	435270	99.9% (-0.1%)
kpvz7ki2wtcwwvo4.onion	Black	130049	130111	100.0% (0.0%)

Table 6: *Test results for various .onion URLS, including the testing phase, script approximation, manual download result and offset. (Given values are in bytes.)*

6 Discussion

For samples from the *surface web*, the search engines only show the first couple of hundred results, thus not allowing to select samples from the full set available. This resulted in a sampling bias. The reason for using Bing, Google and Yahoo is based on the fact that they allow for using a random offset in the query URL. Also, these search engines all show up to 500 results (if available). The reason for selecting 10 pages per pivot word is to create a sufficiently large sample size and mitigate outliers.

The overlap analysis was performed on onion lists, whereas [8] and [11] use this technique for estimating the size of the web by making use of search engines. Whether using the same technique on links gathered via onion lists is sufficiently indicative is questionable. Furthermore, the results for TOR were gathered using HTTP(S), while various other protocols could give access to more resources. The results represented in this paper are therefore a subset of the total size.

The availability of sites on TOR is lower than on the surface web. Several random sample lists were created, after which as many pages were measured as possible. For TOR, 99 samples were used to estimate the mean page size – as opposed to 810 for the surface web. It is unsure whether this amount is sufficient to base the mean page size upon. Also, the samples that were scraped were derived in a breadth-first approach, while the depth of the pages and the average depth remains unknown.

The results for the mean page sizes of both the surface web as well as TOR are very consistent. This means that extending the amount of samples used will not change the mean page size much.

The calculations for both the surface web as well as the deep web accessible through TOR both have a certain margin of error. The final results were rounded to the second decimal.

7 Future work

The initial list of about 47,000 addresses are a good starting point for measuring how deep those domains reach, as well as to measure the average depth of sites on TOR. Not only should further effort be put into researching different parts of TOR (e.g. sites that require authentication) but also the dark web as a whole (I2P, Freenet, etc.). Furthermore, effort could be put into researching the size of TOR over protocols other than HTTP(S), such as (s)FTP.

References

[1] B. He, M. Patel, Z. Zhang, and K. C.-C. Chang, "Accessing the deep web," *Communications of the ACM*, vol. 50, no. 5, pp. 94–101, 2007.

[2] Wikipedia. Surface web. [Online]. Available: https://en.wikipedia.org/wiki/Surface_web

[3] H. Chen, *Dark web: Exploring and data mining the dark side of the web.* Springer Science & Business Media, 2011, vol. 30.

[4] P. G. Ipeirotis, L. Gravano, and M. Sahami, "Probe, count, and classify: categorizing hidden web databases," in *ACM SIGMOD Record*, vol. 30, no. 2. ACM, 2001, pp. 67–78.

[5] Z. Rais. The deep web is 96% of the internet, google know only 4% of it. [Online]. Available: https://medium.com/@zayedrais/the-deep-web-is-96-of-the-internet-google-know-only-4-of-it-819cd53fa7c6

[6] M. Rice. The deep web is the 99% of the internet you can't google. [Online]. Available: https://curiosity.com/topics/the-deep-web-is-the-99-of-the-internet-you-cant-google-curiosity/

[7] E. Staff. The deep web, the dark web & tor: How to browse the secret internet. [Online]. Available: https://www.whoishostingthis.com/blog/2017/03/07/tor-deep-web/

[8] M. K. Bergman, "White paper: the deep web: surfacing hidden value," *Journal of electronic publishing*, vol. 7, no. 1, 2001.

[9] D. Shestakov *et al.*, "Search interfaces on the web: Querying and characterizing," 2008, nA.

[10] A. van den Bosch, T. Bogers, and M. de Kunder, "Estimating search engine index size variability: a 9-year longitudinal study," *Scientometrics*, vol. 107, no. 2, pp. 839–856, May 2016. [Online]. Available: https://doi.org/10.1007/s11192-016-1863-z

[11] K. Bharat and A. Broder, "A technique for measuring the relative size and overlap of public web search engines," *Computer Networks and ISDN systems*, vol. 30, no. 1-7, pp. 379–388, 1998.

[12] S. Gupta and K. K. Bhatia, "A comparative study of hidden web crawlers," *arXiv preprint arXiv:1407.5732*, 2014.

[13] C. Schuijt. Repository containing scripts for scraping the dark web. [Online]. Available: https://github.com/Coen-Schuijt/tor-scraper

[14] de Kunder. The size of the world wide web (the internet). [Online]. Available: https://www.worldwidewebsize.com/

Appendix A URL list

ID	Seed	Web
1	https://thehiddenwiki.org/	Surface
2	https://ahmia.fi/address	Surface
3	https://github.com/alecmuffett/real-world-onion-sites	Surface
4	https://onions.danwin1210.me/?format=text	Surface
5	https://github.com/AgentWotes/onion-links	Surface
6	https://github.com/kenorb/cicada-2014/blob/master/stage11/scripts/onions-list.txt	Surface
7	https://deep-weblinks.com/deep-web-links/	Surface
A	http://visitorfi5kl7q7i.onion/	Tor
B	http://underdj5ziov3ic7.onion/	Tor
C	http://jh32yv5zgayyyts3.onion/	Tor
D	http://wikitjerrta4qgz4.onion/	Tor
E	http://torlinkbgs6aabns.onion/	Tor
F	http://donionsixbjtiohce24abfgsffo2l4tk26qx464zylumgejukfq2vead.onion/	Tor
G	http://torvps7kzis5ujfz.onion/ user/	Tor

Table 7: Seed lists accesses through TOR, including their respective sizes.