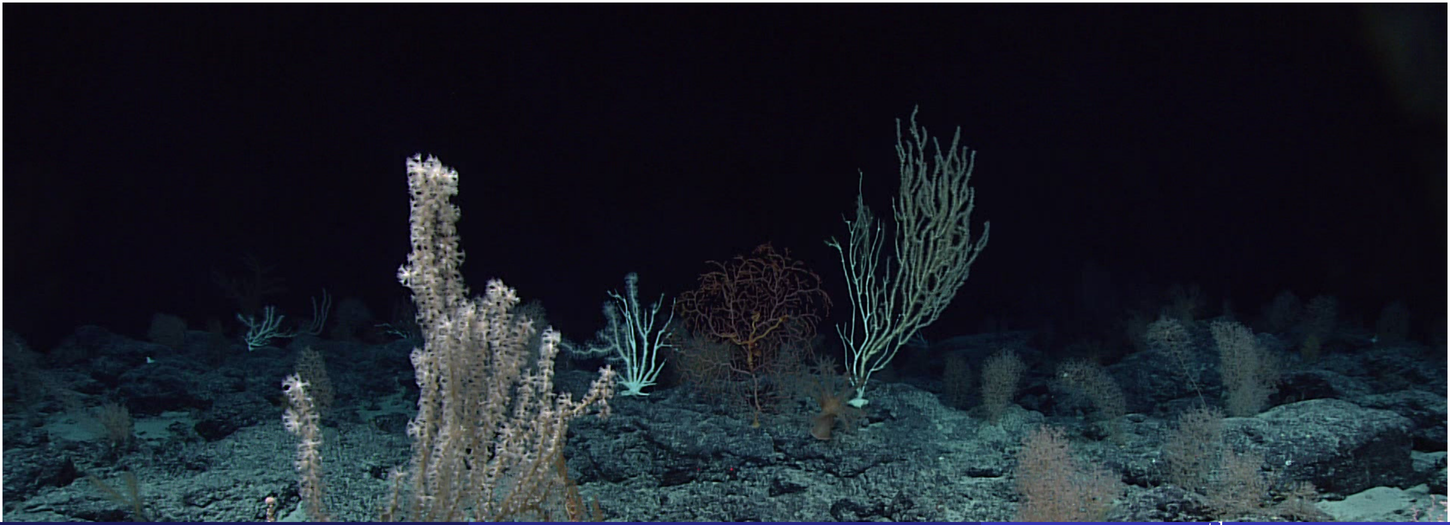
A Deep Dive into the Dark Web

Coen Schuijt

UvA — OS3

February 5th, 2019



Outline

1. Introduction
   - Related work
   - Research Questions
2. Methodologies
   - Surface web
   - TOR
3. Results
4. Conclusion
5. Discussion
6. Future work
7. Questions

Introduction

Figure 1: Graphical overview of the web.

Related work

Surface Web

- M. K. Bergman, “White paper: the deep web: surfacing hidden value,”Journal of electronic publishing, vol. 7, no. 1, 2001
- 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

Deep Web

- S. Raghavan and H. Garcia-Molina, “Crawling the hidden web,” Stanford, Tech. Rep., 2000.
- H. Chen, Dark web: Exploring and data mining the dark side of the web. Springer Science & Business Media, 2011, vol. 30.

Research Questions

The main research question

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

Additional questions

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

Research Questions

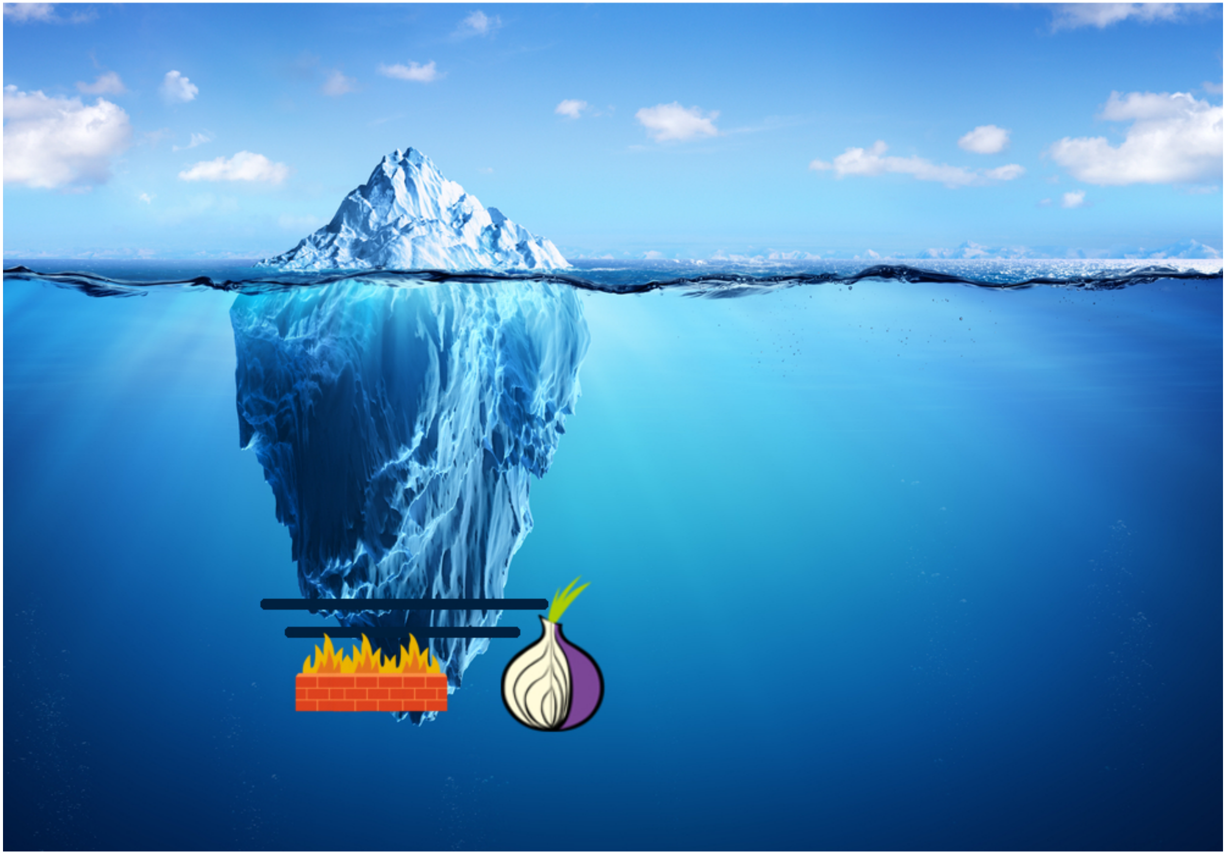


Figure 2: Parts of the web being compared.

Methodologies

Main approach:

1. Amount of pages (surface)
2. Average page size (surface)
3. Amount of pages (TOR)
4. Average page size (TOR)
5. Calculate sizes and ratio

Methodologies: Surface

Amount of pages

- Literature

Page size

- 27 pivot words – several frequency ranks
- 3 search engines
- 10 pages
- $27 \times 3 \times 10 = 810$ samples
- Mean: $\overline{x}(p) = \frac{1}{N} \sum_{i=1}^{N} x_i$
- Deviation (upper lower bounds + confidence interval)

Methodologies: TOR

Amount of pages

- Scrape
- Overlap analysis
- Online source

Page size

- Measure
  - Build
  - Test (white, grey, black)
  - Optimize
- Mean: $\overline{y}(p) = \frac{1}{M} \sum_{i=1}^{M} y_i$
- Deviation (upper lower bounds + confidence interval)

Methodologies: TOR (cont.)

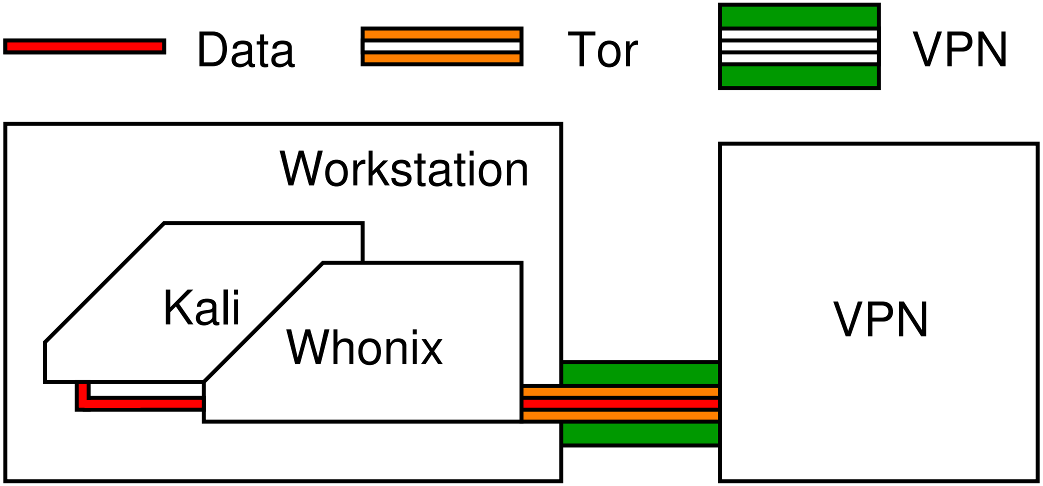


Figure 3: Test setup

Methodologies: TOR (cont.)

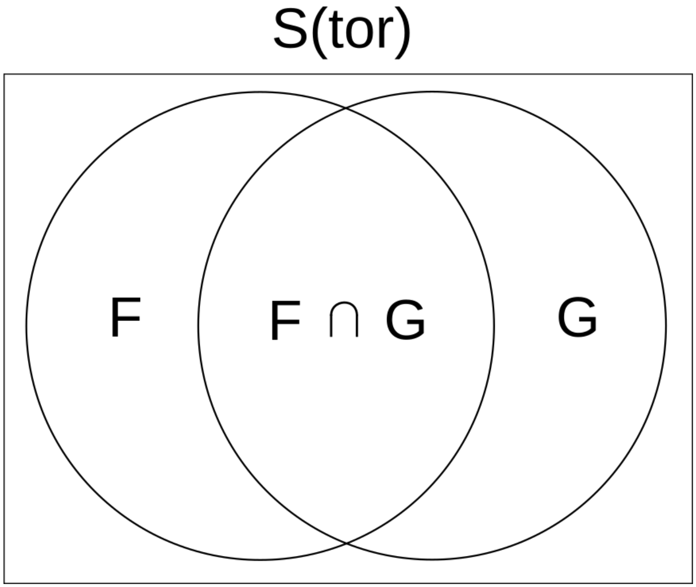


Figure 4: Overlap analysis

Methodologies: TOR (cont.)

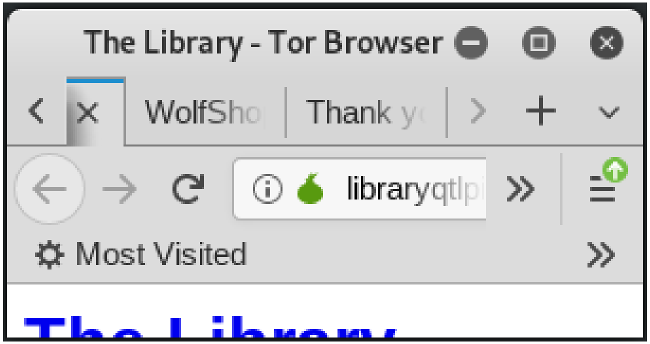


Figure 5: Black box testing

Results: surface

Amount of pages:

- Lower bound [$\mathrm{S}_{\mathrm{L}}(\text{surface})$]: at least 6 billion
- Upper bound [$\mathrm{S}_{\mathrm{U}}(\text{surface})$]: up to 53 billion
- Thursday, January 24th
- Source: `https://www.worldwidewebsize.com/` – (van den Bosch et al.)

Average Page size:

- $\mathrm{N} = 810$
- $\overline{x}(p) = 3483$ KiB
- $\pm$ 529 KiB (CI 95%)
- So
  - Lower bound [$\overline{x}(p_{\mathrm{L}})$]: 2955 KiB
  - Upper bound [$\overline{x}(p_{\mathrm{U}})$]: 4012 KiB

Results: surface (cont.)

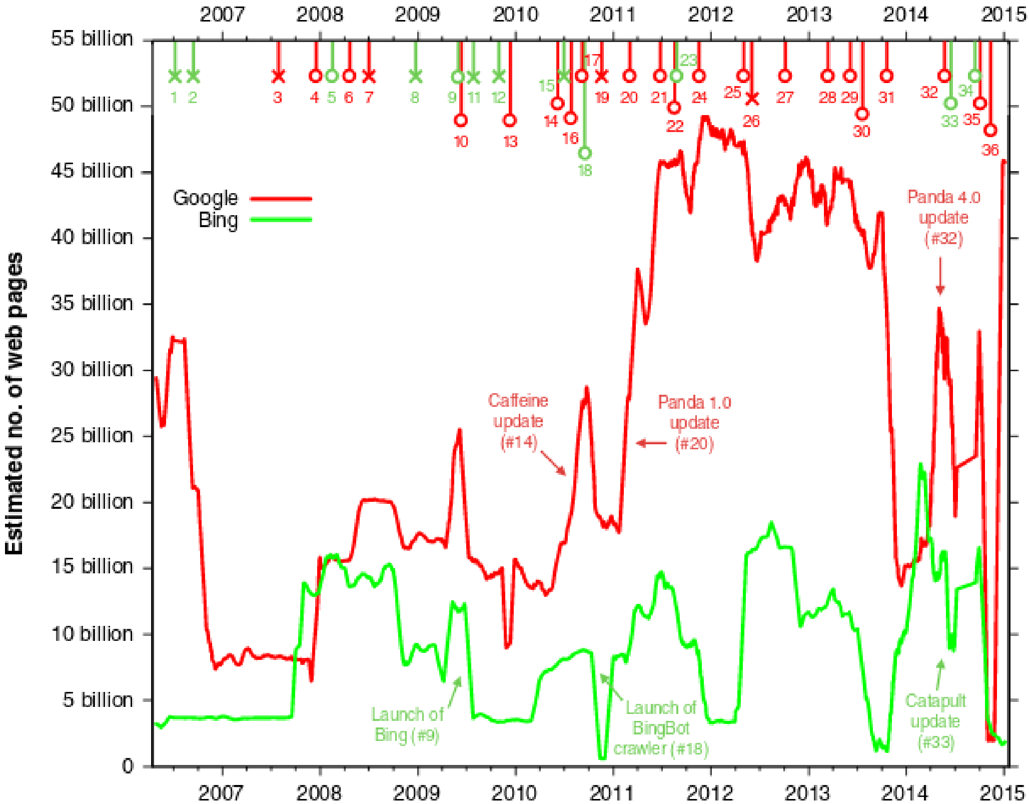


Figure 6: Unweighted averages of 31 days (van den Bosch et al., 2016)

Results: surface (cont.)

Amount of pages:

- Size lower bound [$\mathrm{S_L}$(surface)]: at least 6 billion
- Size upper bound [$\mathrm{S_U}$(surface)]: up to 53 billion
- Thursday, January 24th
- Source: `https://www.worldwidewebsize.com/` – (van den Bosch et al.)

Average Page size:

- N = 810
- $\overline{x}(p)$ = 3483 KiB
- ± 529 KiB (CI 95%)
- So
  - Lower bound [$\overline{x}(p_\mathrm{L})$]: 2955 KiB
  - Upper bound [$\overline{x}(p_\mathrm{U})$]: 4012 KiB

Results: surface (cont.)

Approximate estimations:

Web Size	**Page Size**	**Equation**	**Result**
S_L(surface)	$\overline{x}(p_\text{L})$	$6 \times 10^9 \times \approx 2955$ KiB	≈ 16.12 PiB
S_L(surface)	$\overline{x}(p_\text{U})$	$6 \times 10^9 \times \approx 4012$ KiB	≈ 21.89 PiB
S_U(surface)	$\overline{x}(p_\text{L})$	$53 \times 10^9 \times \approx 2955$ KiB	≈ 142.43 PiB
S_U(surface)	$\overline{x}(p_\text{U})$	$53 \times 10^9 \times \approx 4012$ KiB	≈ 193.40 PiB

Table 1: Size estimations for the surface web

- Reminder: PiB != PB
- 1 PB $= 10^{15}$
- 1 PiB $= 2^{50} (+ \approx 12,6\%)$
  - Total lower bound [$T_\text{L}(surface)$]: 16.12 – 21.89 PiB
  - Total upper bound [$T_\text{U}(surface)$]: 142.43 – 193.40 PiB

Results: TOR

Amount of pages:

- Scraped 46779 pages
- 14 Seed URL's

Results: TOR (cont.)

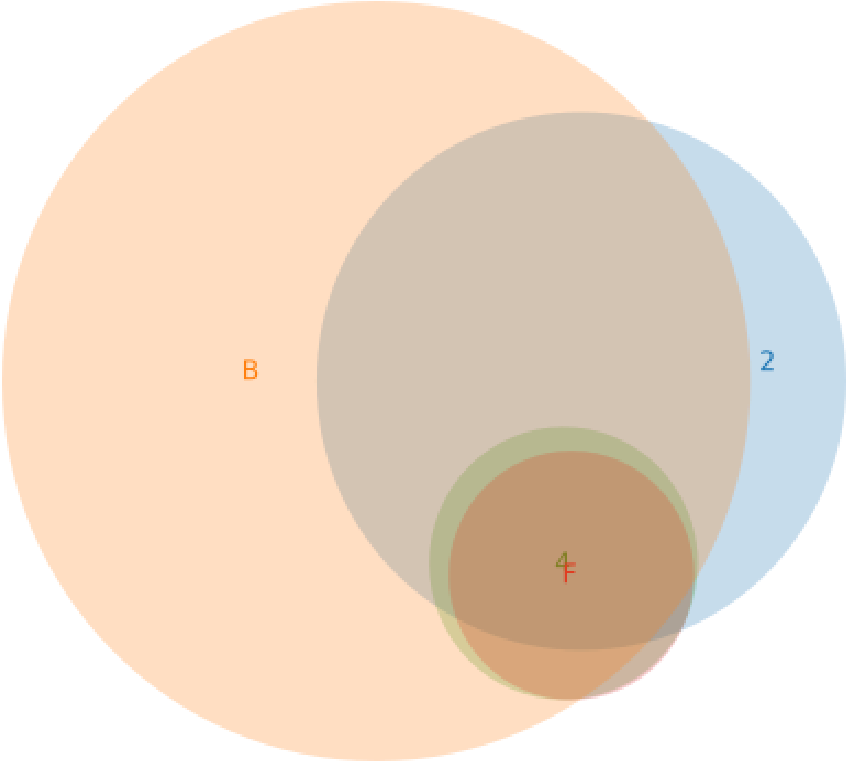


Figure 7: Overlap analysis mixed (numbers for surface, letters for TOR).

Results: TOR (cont.)

Amount of pages:

- Ratio = A ∩ B / B
- $17108/41459 \approx 0.41$

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

Table 2: Estimations of onion web sites, based on overlap of several seed lists.

(2) ahmia.fi

(4) onions.danwin1210.me

(B) underdj5ziov3ic7.onion

(F) donionsixbjtiohve24abfgsffo2l4tk26qx464zylumgejukfq2vead.onion

Results: TOR (cont.)

Amount of pages:

- $\approx$ 50.40K
- Only entry points (breadth first search)
- Average depth of ?
- haystack (`haystakvxad7wbk5.onion`) claims 1.5B pages
- According to `https://onions.danwin1210.me/`:
  - 227/4400 pages > 7days ($\approx$ 5.2%) [January 28th, 2019]
  - 5.2% of 50401 $\approx$ 2600 pages > 7days
  - 50401 - 2600 = 47801 new pages/week
  - 47801 × 52 = 2.485.652 pages/year

Results: TOR (cont.)

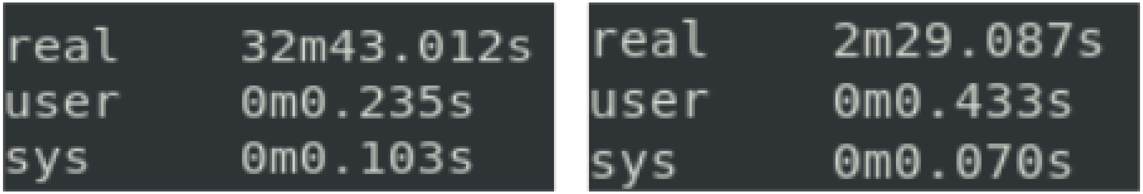
Amount of pages:

- 1.5 billion
- Lower bound $[\mathrm{S_L(tor)}]$: $(1.5 \times 10^9)/0.99 \approx 1.5$ billion sites
- Lower bound $[\mathrm{S_U(tor)}]$: $(1.5 \times 10^9)/0.41 \approx 3.6$ billion sites

Average Page size:

- $\mathrm{N} = 99$
- $\overline{y}(p) = 227$ KiB
- $\pm$ 26 KiB (CI 95%)
- So
  - Lower bound [$\overline{y}(p_{\mathrm{L}})$]: 200 KiB
  - Upper bound [$\overline{y}(p_{\mathrm{U}})$]: 253 KiB

Results: TOR (cont.)



```
real	32m43.012s
user	0m0.235s
sys	0m0.103s
```

```
real	2m29.087s
user	0m0.433s
sys	0m0.070s
```

Figure 8: Timings for synchronous and asynchronous measuring

Results: TOR (cont.)

Approximate estimations:

Web Size	**Page Size**	**Equation**	**Result**
$\mathrm{S}_\mathrm{L}(\mathrm{tor})$	$\overline{y}(p_\mathrm{L})$	$1.5{\times}10^9{\times} \approx 200$ KiB	≈ 0.28 PiB
$\mathrm{S}_\mathrm{L}(\mathrm{tor})$	$\overline{y}(p_\mathrm{U})$	$1.5{\times}10^9{\times} \approx 253$ KiB	≈ 0.35 PiB
$\mathrm{S}_\mathrm{U}(\mathrm{tor})$	$\overline{y}(p_\mathrm{L})$	$3.6{\times}10^9{\times} \approx 200$ KiB	≈ 0.66 PiB
$\mathrm{S}_\mathrm{U}(\mathrm{tor})$	$\overline{y}(p_\mathrm{U})$	$3.6{\times}10^9{\times} \approx 253$ KiB	≈ 0.84 PiB

Table 3: Size estimations for TOR

- Reminder: PiB != PB
- 1 PB $= 10^{15}$
- 1 PiB $= 2^{50}(+ \approx 12,6\%)$
  - Total lower bound [$T_\mathrm{L}(tor)$]: 0.28 – 0.35 PiB
  - Total upper bound [$T_\mathrm{U}(tor)$]: 0.66 – 0.84 PiB

Results: TOR (cont.)

Comparison:

- Surface web: 16.12 – 193.40 PiB (mean 93.46 PiB)
- TOR: 0.27 – 0.35 PiB (mean 0.53)
- $(0.53 / 93.46) \times 100\% \approx 0.6\%$

Conclusion

- About 6 – 53 B pages (surface)
- About 1.5 – 3.6 B pages (TOR)
- Page size 3000 – 4000 KiB (surface)
- Page size 200 – 250 KiB (TOR)
- Surface web is about 93.46 PiB
- TOR accessible is about 0.53 PiB
- TOR is about 0.6% of surface web

Discussion

- Just HTTP ...
- Biases
  - Sampling Bias
  - ...
- Seed lists sufficient?
- Overlap suitable?
- Sample size big enough?
- Moving towards surface?
- ...

Future work

- Gather more data
- Over a longer period
- Extend scraper (depth)
- Other parts (fw, login, etc.)
- Other protocols
- etc.

Questions

Q & A