Outline

1. Introduction
   - Related work
   - Research Questions

2. Methodologies
   - Surface web
   - TOR

3. Results

4. Conclusion

5. Discussion

6. Future work

7. Questions
**Introduction**

![Graphical overview of the web.](image)

**Figure 1:** Graphical overview of the web.
## Related work

### Surface Web

### Deep Web
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: $\bar{x}(p) = \frac{1}{N} \sum_{i=1}^{N} x_i$
- 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)
Figure 3: Test setup
Methodologies: TOR (cont.)

Figure 4: Overlap analysis

S(tor)

F

F ∩ G

G
Methodologies: TOR (cont.)

Figure 5: Black box testing
Amount of pages:

- Lower bound \([ S_L(\text{surface}) ]\): at least 6 billion
- Upper bound \([ S_U(\text{surface}) ]\): up to 53 billion
- Thursday, January 24\(^{th}\)
- Source: https://www.worldwidewebsize.com/ – (van den Bosch et al.)

Average Page size:

- \(N = 810\)
- \(\bar{x}(p) = 3483\) KiB
- \(\pm 529\) KiB (CI 95%)
- So
  - Lower bound \([ \bar{x}(p_L) ]\): 2955 KiB
  - Upper bound \([ \bar{x}(p_U) ]\): 4012 KiB
Figure 6: Unweighted averages of 31 days (van den Bosch et al., 2016)
Results: surface (cont.)

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

Average Page size:
- $N = 810$
- $\bar{x}(p) = 3483$ KiB
- $\pm 529$ KiB (CI 95%)
- So
  - Lower bound $[\bar{x}(p_L) ]$: 2955 KiB
  - Upper bound $[\bar{x}(p_U) ]$: 4012 KiB
Approximate estimations:

<table>
<thead>
<tr>
<th>Web Size</th>
<th>Page Size</th>
<th>Equation</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_L$ (surface)</td>
<td>$\bar{x}(p_L)$</td>
<td>$6 \times 10^9 \times \approx 2955 \text{ KiB}$</td>
<td>$\approx 16.12 \text{ PiB}$</td>
</tr>
<tr>
<td>$S_L$ (surface)</td>
<td>$\bar{x}(p_U)$</td>
<td>$6 \times 10^9 \times \approx 4012 \text{ KiB}$</td>
<td>$\approx 21.89 \text{ PiB}$</td>
</tr>
<tr>
<td>$S_U$ (surface)</td>
<td>$\bar{x}(p_L)$</td>
<td>$53 \times 10^9 \times \approx 2955 \text{ KiB}$</td>
<td>$\approx 142.43 \text{ PiB}$</td>
</tr>
<tr>
<td>$S_U$ (surface)</td>
<td>$\bar{x}(p_U)$</td>
<td>$53 \times 10^9 \times \approx 4012 \text{ KiB}$</td>
<td>$\approx 193.40 \text{ PiB}$</td>
</tr>
</tbody>
</table>

Table 1: Size estimations for the surface web

- Reminder: PiB $\neq$ PB
- 1 PB $= 10^{15}$
- 1 PiB $= 2^{50} (+ \approx 12, 6\%)$
  - Total lower bound $[ T_L(surface) ]$: 16.12 – 21.89 PiB
  - Total upper bound $[ T_U(surface) ]$: 142.43 – 193.40 PiB
Amount of pages:

- Scraped 46,779 pages
- 14 Seed URL’s
Figure 7: Overlap analysis mixed (numbers for surface, letters for TOR).
Amount of pages:

- Ratio = \( A \cap B / B \)
- \( 17108/41459 \approx 0.41 \)

<table>
<thead>
<tr>
<th>A</th>
<th>A</th>
<th>B</th>
<th>B</th>
<th>A ( \cap ) B</th>
<th>Ratio</th>
<th>Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>20798</td>
<td>B</td>
<td>41459</td>
<td>17108</td>
<td>0.41</td>
<td>20798/0.41 = 50401</td>
</tr>
<tr>
<td>2</td>
<td>20798</td>
<td>4</td>
<td>5352</td>
<td>4511</td>
<td>0.84</td>
<td>20798/0.84 = 24675</td>
</tr>
<tr>
<td>2</td>
<td>20798</td>
<td>F</td>
<td>4461</td>
<td>3700</td>
<td>0.83</td>
<td>20798/0.83 = 25075</td>
</tr>
<tr>
<td>B</td>
<td>41459</td>
<td>4</td>
<td>5352</td>
<td>5143</td>
<td>0.96</td>
<td>41459/0.96 = 43143</td>
</tr>
<tr>
<td>B</td>
<td>41459</td>
<td>F</td>
<td>4461</td>
<td>4250</td>
<td>0.95</td>
<td>41459/0.95 = 43517</td>
</tr>
<tr>
<td>4</td>
<td>4461</td>
<td>F</td>
<td>4461</td>
<td>4423</td>
<td>0.99</td>
<td>4461/0.99 = 4499</td>
</tr>
</tbody>
</table>

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

(2) ahmia.fi

(4) onions.danwin1210.me

(B) underdj5ziow3ic7.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 28\(^{th}\), 2019]
  - 5.2% of 50401 \( \approx 2600 \) pages > 7days
  - 50401 - 2600 = 47801 new pages/week
  - 47801 \times 52 = 2.485.652 \) pages/year
Amount of pages:

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

Average Page size:

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

Figure 8: Timings for synchronous and asynchronous measuring
Approximate estimations:

<table>
<thead>
<tr>
<th>Web Size</th>
<th>Page Size</th>
<th>Equation</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_L$(tor)</td>
<td>$\bar{y}(p_L)$</td>
<td>$1.5 \times 10^9 \times \approx 200 \text{ KiB}$</td>
<td>$\approx 0.28 \text{ PiB}$</td>
</tr>
<tr>
<td>$S_L$(tor)</td>
<td>$\bar{y}(p_U)$</td>
<td>$1.5 \times 10^9 \times \approx 253 \text{ KiB}$</td>
<td>$\approx 0.35 \text{ PiB}$</td>
</tr>
<tr>
<td>$S_U$(tor)</td>
<td>$\bar{y}(p_L)$</td>
<td>$3.6 \times 10^9 \times \approx 200 \text{ KiB}$</td>
<td>$\approx 0.66 \text{ PiB}$</td>
</tr>
<tr>
<td>$S_U$(tor)</td>
<td>$\bar{y}(p_U)$</td>
<td>$3.6 \times 10^9 \times \approx 253 \text{ KiB}$</td>
<td>$\approx 0.84 \text{ PiB}$</td>
</tr>
</tbody>
</table>

Table 3: Size estimations for TOR

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

- Surface web: 16.12 – 193.40 PiB (mean 93.46 PiB)
- TOR: 0.27 – 0.35 PiB (mean 0.53)
- \( \left( \frac{0.53}{93.46} \right) \times 100\% \approx 0.6\% \)
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?
- ...

Coen Schuijt (UvA — OS3) A Deep Dive into the Dark Web February 5th, 2019 26 / 28
Future work

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