# Spiders like Onions: on the Network of Tor Hidden Services

Massimo Bernaschi Alessandro Celestini Stefano Guarino Flavio Lombardi Enrico Mastrostefano

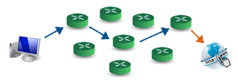
Institute for Applied Computing, National Research Council of Italy

## Tor: Network vs. Web vs. Web Graph

The **Tor network** is an overlay network composed of thousands of routers, called relays. It is an **anonymity** network, representing one of the main infrastructures of the **DarkWeb**.

The Tor network allows:

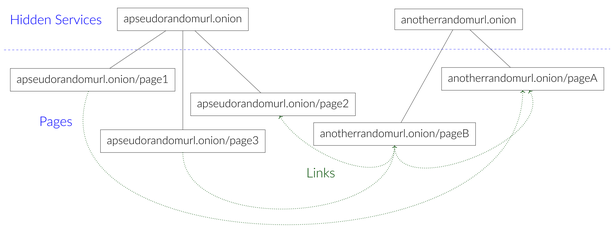
- accessing the **Internet** anonymously;
- running **anonymous Web servers**, known as hidden services.




Hidden services:

- can only be accessed using a **Tor-enabled browser**;
- are identified by a non-mnemonic **16-character string**;
- have a specific **.onion** domain.

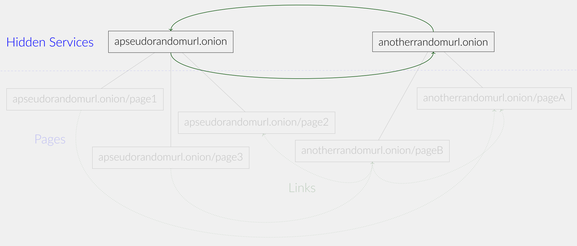
Web resources hosted on Tor constitute the **Tor Web**. Tor Web pages may be connected through **hyperlinks**, thus defining a **Web graph**.




We study the properties of the Tor Web Graph aggregated by Hidden Service.

## The Tor Hidden Services Graph

Vertex set: a vertex for each Tor hidden service found
Edge set: an edge $A \to B$ if there exists at least one link from a page of service A to a page of service B



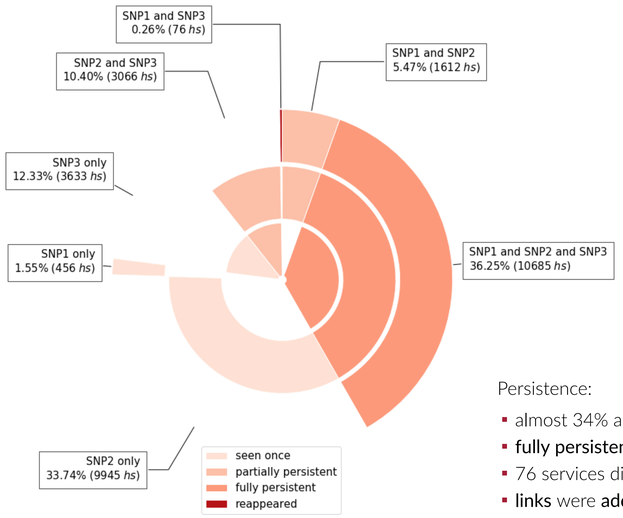

Previously @CRANIC (IAC-CNR):

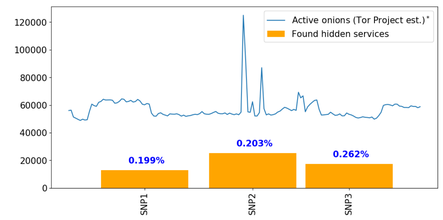
[BCGL17] Massimo Bernaschi, Alessandro Celestini, Stefano Guarino, and Flavio Lombardi, *Exploring and Analyzing the Tor Hidden Services Graph*, ACM Trans. Web **11** (2017), no. 4, 24:1–24:26.

[CG17] Alessandro Celestini and Stefano Guarino, *Design, implementation and test of a flexible tor-oriented web mining toolkit*, Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics (Amantea, Italy), WIMS '17, ACM, 2017, pp. 19:1–19:10.

## Data, Graphs, Persistence

We scraped the Tor Web for five months, obtaining **three snapshots** of the Tor Web which constitute the **widest exploration** of the Tor Web in the literature. To capture persistent features, we defined a **CORE** Graph, which is induced by **stable edges**.




* This includes:
(i) hidden services that do not host websites;
(ii) onion urls only created for accessing Tor-specific messaging services.

| snapshot | # pages contacted (by response type) 2xx | 3xx | 4xx | 5xx | Total |
|---|---|---|---|---|---|
| SNP1 | 1821842 | 277813 | 197128 | 141205 | **2437989** |
| SNP2 | 2339718 | 471519 | 262403 | 324552 | **3398192** |
| SNP3 | 765876 | 393018 | 105406 | 67115 | **1331415** |

3xx: Web redirection; 4xx and 5xx: error codes.

Persistence:

- almost 34% appeared in **SNP2 only**;
- **fully persistent** hidden services are $< 36\%$;
- 76 services disappeared after SNP1 to **reappear** in SNP3;
- **links** were **added/removed** even among persistent services.

## Graph Structure

**Web** graphs are typically **scale-free** (both in- and out-degrees are power-laws) and have a **bow-tie** structure. In all Tor Hidden Services Graphs (SNP1, SNP2, SNP3, CORE), the **out-degree** is **not** a power-law, it has **heavier bulk** and **tail**. The Graphs have a **small average distance** and a **pronounced clustering/transitivity**. **No** simple graph **model** fully explains this structure.

**Erdos-Renyi**

$C \sim \langle \deg \rangle / |V|$
$\langle dist \rangle \sim \ln(|V|)$

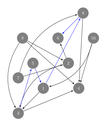
**Scale-Free**

$\lambda > 3 : \langle dist \rangle \sim \ln(|V|)$
$2 < \lambda < 3 : \langle dist \rangle \sim \ln \ln(|V|)$

**Small-World**

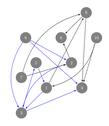
$C \gg \langle \deg \rangle / |V|$
$\langle dist \rangle \sim \ln(|V|)$

**Paths length**



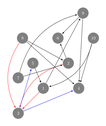
| graph | $\langle dist \rangle$ | $\ln \lvert V \rvert$ | $\ln \ln \lvert V \rvert$ |
|---|---|---|---|
| SNP1 | 3.79316 | 9.459 | 2.247 |
| SNP2 | 4.96028 | 10.138 | 2.316 |
| SNP3 | 3.66455 | 9.767 | 2.279 |
| CORE | 3.98291 | 8.944 | 2.190 |

**Clustering/Transitivity**



| $C$ | $T_1$ | $T_2$ | $2\langle \deg_{in} \rangle / \lvert V \rvert$ |
|---|---|---|---|
| 0.00943 | 0.00617 | 0.0039 | 0.000877 |
| 0.00407 | 0.00779 | 0.0016 | 0.000352 |
| 0.00492 | 0.00253 | 0.00197 | 0.000678 |
| 0.00873 | 0.00535 | 0.00356 | 0.000983 |

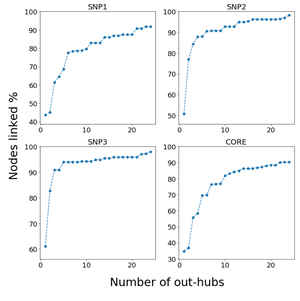
**Degree distribution**



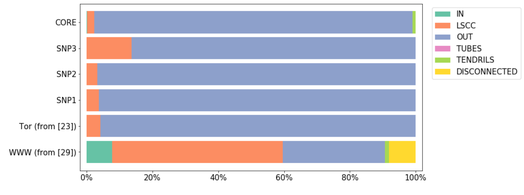
| $\deg_{in}$ | $\deg_{out}$ |
|---|---|
| $\lambda = -2.88$ | heavy-tail |
| $\lambda = -3.32$ | heavy-tail |
| $\lambda = -3.35$ | heavy-tail |
| $\lambda = -2.7$ | heavy-tail |

The Graph is dominated by few **out-hubs** (mostly hidden directories) that link to most of the other hidden services. This also emerges in its highly **asymmetrical** bow-tie structure, which is always single-ended except for the **CORE** Graph.

- 83% to 95% of the services are **sinks**;
- Top **out-hubs** are $50\times$ greater than top in-hubs;
- The 10 top out-hubs link to 80% to 95% of the **entire** graph.



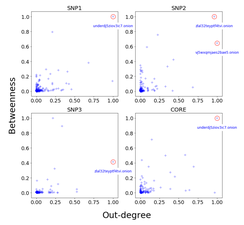

- In all snapshots, the LSCC is **small**, the OUT set is **very large**, all other sets are **empty**;
- The **CORE** is similar, but the IN is $\approx 0.1\%$ and the TENDRILS are $\approx 1\%$.




## Centralities and Central Hidden Services

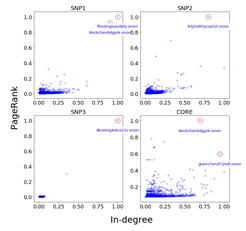
There is **no** evident **correlation** among centrality metrics in the Tor Hidden Services Graph. Yet, **important** nodes have either:

Large **Out-degree & Betweenness**




| onion | Topic |
|---|---|
| underdj5ziov3ic7 | Directory |
| zlal32teyptf4tvi | Directory |
| vj5wxqmjaes2bae5 | Directory |

Large **In-degree & PageRank**




| onion | Topic |
|---|---|
| fhostingesps6bly | Hidden Wiki |
| blockchainbdgpzk | Cryptocurrency |
| tt3j2x4k5ycaa5zt | Directory/Hosting |
| dhosting4okcs22v | Hosting |
| grams7enufi7jmdl | Market SE |

After **removing** the 10 top out-hubs from the CORE:

- $C$ **grew** but $\langle dist \rangle$ remained almost **unvaried**;
- the **PageRank** can be used to identify **markets**, while the **Betweenness** becomes **less informative**.

Betweenness

| onion | Topic |
|---|---|
| auutwvpt2zktxwng | Directory |
| zqktlwi4fecvo6ri | Hidden Wiki |
| deepdot35wvmeyd5 | Market |
| torpress2sarn7xw | News/Blogs |
| torvps7kzis5ujfz | News/Blogs |

PageRank

| onion | Topic |
|---|---|
| grams7enufi7jmdl | Market SE |
| deepdot35wvmeyd5 | Market |
| lchudifyeqm4ldjj | Market |
| zqktlwi4fecvo6ri | Hidden Wiki |
| tmskhzavkycdupbr | Market |

## Conclusions

Analyzing Tor over time:

- Exploring **all** Tor is **virtually impossible** (with a crawler).
- Tor services are **highly volatile**, but there is a significant **stable** core, with slightly different properties.
- **Connections** are also volatile (possibly due to **volatile pages**).

Studying the topology of the hidden services graph:

- The Tor Web graph is (ultra?) **small**, only **partially** governed by **preferential attachment**, with clustering **not** large enough for being a **(WS) small-world**.
- The **out-hubs determine** the topology and may **hide** other features.
- Central services are mostly **link directories** and **markets**.

Find more on **http://www.cranic.it**