Spiders like Onions: on the Network of Tor Hidden Services

Massimo Bernaschi
Institute for Applied Computing,
National Research Council of Italy
Rome, Italy
massimo.bernaschi@cnr.it

Alessandro Celestini
Institute for Applied Computing,
National Research Council of Italy
Rome, Italy
a.celestini@iac.cnr.it

Stefano Guarino
Institute for Applied Computing,
National Research Council of Italy
Rome, Italy
s.guarino@iac.cnr.it

Flavio Lombardi
Institute for Applied Computing,
National Research Council of Italy
Rome, Italy
flavio.lombardi@cnr.it

Enrico Mastrostefano
Institute for Applied Computing,
National Research Council of Italy
Rome, Italy
e.mastrostefano@iac.cnr.it

ABSTRACT

Tor hidden services allow offering and accessing various Internet resources while guaranteeing a high degree of provider and user anonymity. So far, most research work on the Tor network aimed at discovering protocol vulnerabilities to de-anonymize users and services. Other work aimed at estimating the number of available hidden services and classifying them. Something that still remains largely unknown is the structure of the graph defined by the network of Tor services. In this paper, we describe the topology of the Tor graph (aggregated at the hidden service level) measuring both global and local properties by means of well-known metrics. We consider three different snapshots obtained by extensively crawling Tor three times over a 5 months time frame. We separately study these three graphs and their shared "stable" core. In doing so, other than assessing the renowned volatility of Tor hidden services, we make it possible to distinguish time dependent and structural aspects of the Tor graph. Our findings show that, among other things, the graph of Tor hidden services presents some of the characteristics of social and surface web graphs, along with a few unique peculiarities, such as a very high percentage of nodes having no outbound links.

CCS CONCEPTS

• **Information systems** → **Web mining**; **Deep web**; *Information retrieval*; • **Mathematics of computing** → *Graph theory*.

KEYWORDS

Tor; Web Graph; Dark Web; Complex Networks

ACM Reference Format:
Massimo Bernaschi, Alessandro Celestini, Stefano Guarino, Flavio Lombardi, and Enrico Mastrostefano. 2019. Spiders like Onions: on the Network of Tor Hidden Services. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3308558.3313687

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313687

1 INTRODUCTION

Research efforts on The Onion Router network (Tor) have recently flourished, focusing on evaluating its security [8], understanding its evolution [24], and discussing its thematic organization [38]. Nevertheless, the limited number of entry points to Tor on the surface Web makes it difficult to fully uncover many of Tor's characteristics. In particular, despite some insights provided over the last years [6, 15, 23, 36], few information are available regards the topology and the volatility of the Tor network. In this study we exploit three sets of crawling data collected over three different time frames to present a thorough analysis of the structural properties of the graph of Tor Hidden Services (HS) that makes it possible to tell apart persistent from variable characteristics.

Similarly to the surface WWW [12], the structure of the Tor HS Web graph can be seen as an indicator of intrinsic characteristics of the Tor network and of latent patterns of interactions among Tor users. Since the work of Watts & Strogats [40] and Barabasi & Albert [5] it has been widely recognized that the in-depth study of the properties of the underlying graph is crucial for determining behavioral and structural aspects of a complex system, and for understanding and possibly explaining the emergence of specific features in real world networks. Our efforts for identifying the distinguishing traits of the topology of the Tor Web can therefore be of great help to shed light on the usage patterns, the dynamics and the vulnerabilities of the Tor network.

The paper also addresses the evolution of the Tor Web graph showing the actual changes that take place in the quality and quantity of available services and in the persistence of their interconnections over time. In particular, we provide a rich set of results and discussions on deltas over time that allow for detailed reasoning on Tor Web connection/topological trends. To the best of our knowledge there are no similar studies on the Tor Web. Our present study, albeit limited as regards its timeframe, therefore provides useful information and hints to foster further research in the area.

The rest of the paper is organized as follows: Section 2 reviews background information and related work; Section 3 describes the methodology used for collecting data, extracting the graphs and studying their properties; Section 4 analyzes the obtained graphs and details on results; Section 5 discusses in depth our findings, in comparison with well-known graph models; finally, Section 6 draws conclusions and suggests directions for future work.

2 BACKGROUND AND RELATED WORK

Past research work on Tor has mainly been devoted to assess its vulnerabilities. However, as a positive side effect, novel data and insights on Tor services and network have been obtained.

Biryukov *et. al.* [8] in 2013 exploited a Tor vulnerability to collect all hidden service descriptors in approximately 2 days using a modest amount of resources. They found out that, while the contents of Tor hidden services is rather varied, the most popular hidden services were related to botnets. It is worth noticing that their approach cannot be reproduced, because they exploited a Tor bug that was fixed in recent versions of the software.

In a previous work [6], we leveraged automated Tor network exploration to the purpose of relating semantic content similarity with Tor topology at the page, host, and service level. The present work largely extends on both data collection and analysis over [6].

Also Ghosh *et. al.* [22] developed an automated tool to explore the Tor network and analyze the contents of onion sites. Their classification framework maps onion site content to a set of categories, and clusters services to categorize onion content. The main difference with respect to our work is that they focus on page content/semantics, and do not consider network topology. Owen *et. al.* [35], by operating 40 relays over a 6 month time frame, reported over hidden services persistence, contents, and popularity. Their aim was classifying services based on their content.

Similarly to our present work, Christin *et. al.* [15] collected crawling data on Tor hidden services over an 8 month lifespan. They evaluated the evolution/persistence of such services over time, and performed a study on the content and topology of the explored network. The main difference with our present work is that the Tor graph we explore is much larger, not being limited to a single marketplace. In addition, we present here a more in depth evaluation of the graph topology. De Domenico *et. al.* [17], used the data collected in [4] to study the topology of the Tor network . They gave a characterization of the topology of the Darknet and proposed a generative model for the Tor network to study its resilience. Their viewpoint is quite different from our own here, as they consider the network at the autonomous system (AS) level.

Very recently Griffith *et. al.* [23] performed a topological analysis of the Tor hidden services graph. They crawled the Tor network using the commercial service *scrapinghub.com*, through the *tor2web* proxy onion link. Interestingly, they reported that more than the 87% of Darkweb sites never link to another site. The main difference with our present work lies in both the extent of the explored network (we collected a much more extensive dataset than that accessible through *tor2web*) and the depth of the analysis of the network itself (we evaluate a larger set of network characteristics).

Differently from the Literature on the Tor network, the topology of the WWW has been the subject of a large number of studies in the past. In this paper we also aim at comparing Tor network characteristics with those of the surface Web, briefly surveyed below.

Among random graph models suitable for the surface Web [28],[10] Kleinberg in particular [26] introduced algorithms for improved Web search and automatic community discovery, thus providing one of the first publicly known portraits of the Web graph. Kleinberg also stressed that traditional random graph models, such as the well-known Erdős-Rényi model [19] do not exhibit many properties of the Web graph. Among other studies on the WWW, Broder *et. al.* [12] discovered the heavy-tailed distribution of node degrees, claiming a power-law distribution, and the presence of large hubs along with a peculiar structure of the graph they called *bow-tie*. Bearing in mind that the adopted crawling technique affects the structure of the results, subsequent studies have somewhat reached different results, [1, 18, 37]. In a recent work Meusel *et. al.* [30] analyzed the structure of the WWW at different levels: pages, hosts and pay-level domain. The last aggregation level can be seen as the Tor service level and we will discuss similarities and differences of their results compared to our findings on the Tor HS graph.

3 METHODOLOGY

In this Section we describe the methodology used for collecting data, building the graphs and studying their properties.

3.1 Data Collection

We aim at characterizing the portion of the Tor Web that can be accessed by using a custom web scraping procedure. Specifically, we assembled a large root set by merging onion urls advertised on well-known Tor wikis and link directories (*e.g.*, "The Hidden Wiki" [1]), or obtained from standard (*e.g.*, Google) and Tor-specific (*e.g.*, Ahmia) search engines. Then, in the 5-month time frame between January 2017 and May 2017, we launched our customized crawler three times and let each execution run for about six weeks. Thus, we obtained three different "snapshots" of the Tor Web, denoted SNP1, SNP2, and SNP3, respectively.

The numbers of our datasets, reported in Table 1, are comparable to – and, as a matter of fact, greater than – similar studies in the Literature [7, 23, 38]. Yet, if we refer to the statistics provided by the Tor Project for the corresponding time window [2], our crawls only reached 25% to 35% of the total number of daily published hidden services. It is not clear to which extent those estimates are inflated by the existence of Tor-specific messaging services in which each user is identified by a unique onion domain [23] and by hidden services that do not host websites. In any case, reaching all active onion urls is not arguably possible with ordinary resources [3] and, to the best of our knowledge, the present study is the widest exploration of the Tor Web performed so far.

To access the Tor network and to collect data from hidden services we evaluated different crawlers. In particular, we evaluated the following alternatives: Apache Nutch [4] [25], Heritrix [5] [31] and BUbiNG [9]. By considering criteria such as performance, configurability and extensibility, we found BUbiNG to be the most appropriate choice for our goals. BUbiNG is a high-performance, scalable, distributed, open-source crawler, written in Java, and developed by the Laboratory for Web Algorithmics (LAW) part of the Computer Science Department of the University of Milan. To allow

[1] wikitjerrta4qgz4.onion
[2] https://metrics.torproject.org/hidserv-dir-onions-seen.html?start=2017-01-01&end=2017-05-01
[3] Tor's working principles make it possible to run a hidden service whose existence is only known to the relays where the introductory points of that service are published [34].
[4] http://nutch.apache.org
[5] https://webarchive.jira.com/wiki/display/Heritrix

Table 1: Outcomes of the three crawling processes.

Crawl	End Date	# records per response type				
		2xx	3xx	4xx	5xx	Total
SNP1	22/02/17	1821842	277813	197128	141205	2437989
SNP2	10/04/17	2339718	471519	262403	324552	3398192
SNP3	22/05/17	765876	393018	105406	67115	1331415

A status code 3xx is related to Web redirection (https://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html). Status codes 4xx and 5xx are error codes.

BUbiNG to operate in the Tor network (instead of the surface Web), we used a HTTP Proxy configured with the SOCKS Proxy provided by Tor. After testing some alternatives we chose privoxy [6]. In particular, we decided not to use polipo [7] that is often used in combination with Tor, because it is no longer maintained and seemed unable to correctly manage the format of some HTTP responses. During the crawling phase we observed that some hidden services check the user-agent of the requester and, if it does not match the last version of the Tor Web browser, they reply with an error. This behavior had to be taken into account when collecting data, to allow the crawler to reach the largest possible portion of hidden services. Another issue that raised during the crawling is the load of the Tor client, *i.e.*, the software used to access Tor. We noticed that under stress (*i.e.*, when too many requests are performed in parallel), the Tor client, quite often, does not respond correctly, *i.e.*, it may mistakenly report that a hidden service is not available, even if the service is actually up and running. The maximum load depends on the specifications of the machine where the software runs, and we assessed it for our configuration during the experimental phase.

3.2 Graph Extraction

For each one of the three snapshots we extracted the associated directed Tor Service Graph (SG) aggregating pages at the service level. To extract the graph we used the Graph Builder module of the toolkit presented in [14]. In the SG, each node represents the set of pages belonging to a hidden service, *i.e.*, a Tor domain identified by a sequence of 16 characters (base32 encoded).[8] In the SG an edge connecting one hidden service to another represents the existence of at least one page of the first service that contains a hypertextual link to any page of the second service. Since we only analyzed onion links, all surface web services and all edges from/to the surface web have been ignored and have not been included in the graphs. We believe including surface web nodes/links would have introduced a bias in the Tor network analysis, due to the large difference in scale between the two underlying graphs. Even just introducing surface border nodes would have affected the analysis and it would not have added relevant information.

[6] https://www.privoxy.org
[7] https://www.irif.fr/ jch/software/polipo/
[8] In October/November 2017 a new generation of hidden services was introduced and supported by the Tor browser. They are identified by character sequences of length 56 instead of the usual 16 (https://blog.torproject.org/tors-fall-harvest-next-generation-onion-services). This change was introduced after our data collection period ended in May 2017.

To further clarify how we built the graphs, let us consider the following example, depicted in Figure 1. We found the hidden service:

`duskgytldkxiuqc6.onion`

to host three pages:

`duskgytldkxiuqc6.onion/comsense.html`;
`duskgytldkxiuqc6.onion/fedpapers/`;
`duskgytldkxiuqc6.onion/fedpapers/federa02.htm`.

To represent these resources we use a single node in the SG.

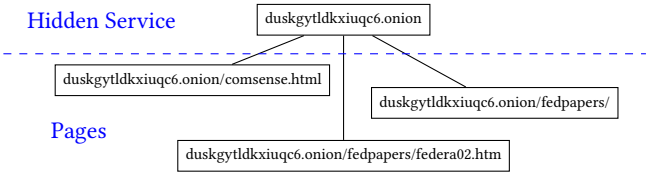


Figure 1: An example of graph construction

Besides the SGs of the three snapshots acquired with our crawls, we considered a fourth graph representing Tor's "stable core". It corresponds to the communal subgraph of SNP1, SNP2 and SNP3 induced by the edges that appear in all the three graphs.

3.3 Graph Analysis

As a first step towards the understanding of Tor dynamics, we compare macroscopic features of the four graphs to assess the persistence of Tor hidden services and of their connections. Next, we characterize the four graphs on both a global and on a local scale. Specifically, for each graph:

- We compute a set of global metrics, including measures of centralization, transitivity and efficiency.
- We extract the in- and out-degree distribution and assess whether these distributions follow a power law.
- We count the number of strongly connected components and characterize the giant strongly connected component (LSCC in the following).
- We consider several centrality measures, draw their distribution and match them with one another.
- We provide and analyze a bow-tie decomposition of the graphs under study.

Based on all gained pieces of information, we infer the general structure of the four graphs, spotting differences and highlighting common aspects that may be assumed to define the topology of Tor hidden services. Additionally, we identify a small set of hidden services that seem to play an especially important role in the graph and, through direct examination, we aimed at explaining why. All symbols and metrics used in the paper are summarized in Table 2.

4 RESULTS

In this Section we report and briefly comment the results of our analysis.

4.1 Services Persistence

As showed by other studies [7, 8, 35], there is a huge variability in the persistence of Tor hidden services. This must be carefully taken

Table 2: Notations and definitions used throughout the present paper.

Symbol	Definition
$G = (V, E)$	Graph with vertex set V and edge set E
N	Number of nodes: $N = \|V\|$
M	Number of edges: $M = \|E\|$
D	Density: $\frac{M}{N(N-1)}$
δ_{in}	Minimum in-degree
Δ_{in}	Maximum in-degree
δ_{out}	Minimum out-degree
Δ_{out}	Maximum out-degree
$\langle \deg \rangle$	Average in/out-degree
ρ	Assortativity: see (26) in [33]
C_{in}	In-degree centralization: $\frac{N*\Delta_{in}-\sum_{v\in V}\deg_{in}(v)}{(N-1)^2}$
C_{out}	Out-degree centralization: $\frac{N*\Delta_{out}-\sum_{v\in V}\deg_{out}(v)}{(N-1)^2}$
C	Global clustering coefficient: $\frac{\#\text{ closed triplets}}{\#\text{ all triplets}}$
T_1	Global transitivity of type 1: $\frac{\#(u,v,w)\text{: } u\to v\wedge v\to w\wedge w\to u}{\#(u,v,w)\text{: } u\to v\wedge v\to w}$
T_2	Global transitivity of type 2: $\frac{\#(u,v,w)\text{: } u\to v\wedge u\to w\wedge (v\to w\vee w\to v)}{\#(u,v,w)\text{: } u\to v\wedge u\to w}$
σ_{vu}	Number of shortest paths from v to u
$\sigma_{vu}(t)$	Number of shortest paths from v to u including t
$dist(v,u)$	Shortest path length from v to u
d	Diameter: $\max_{v\in V}\max_{u\in V} dist(v,u)$
r_{in}	Radius with in-paths: $\min_{v\in V}\max_{u\in V} dist(u,v)$
r_{out}	Radius with out-paths: $\min_{v\in V}\max_{u\in V} dist(v,u)$
$\langle dist \rangle$	Average shortest path length
E_{in}	Global efficiency with in-paths: $\frac{1}{N(N-1)}\sum_{u<v\in V}\frac{1}{d(v,u)}$
E_{out}	Global efficiency with out-paths: $\frac{1}{N(N-1)}\sum_{u<v\in V}\frac{1}{d(u,v)}$
B	Betweenness centrality: $B(v) = \sum_{s\neq v\neq t\in V}\frac{\sigma_{st}(v)}{\sigma_{st}}$
PR	PageRank: see [21]

into consideration in any attempt of characterizing the topology of the Tor Web graph, because by scraping the Tor network we only obtain a snapshot of the hidden services that were active at the time the crawler issued a connection request. More generally, even for the surface Web there is evidence that the crawling process may affect the structure of the extracted graph, leading to incomplete or wrong conclusions [29]. We therefore repeated the entire data collection phase three times over five months in an effort to reduce variance and guarantee the consistency of our results. As a side benefit, we are able to further assess the renowned volatility of Tor hidden services, other than possibly telling apart time dependent from structural features of the Tor web graph.

As a starting point, Figure 2 shows how the total body of accessed hidden services is distributed across the three snapshots. SNP2 clearly emerges as the largest dataset, but this is not surprising if we compare our results with the statistics provided by the Tor Project [9], that show a spike of published hidden services in the second half of March 2017. We note that 10685 hidden services were found by all three crawling runs, suggesting that these services were durably present over the considered five month time frame. This core set represents the 83.3% of services reached during SNP1, the 42.2% of services reached during SNP2 and the 61.2% of services reached during SNP3. We also see that all pairwise intersections are not empty, meaning that during our data harvesting process (at least) 1612 hidden services disappeared, 3066 new hidden services appeared, and most notably, 76 hidden services reappeared after having gone inactive at some point in time. There is also the possibility that some of the hidden services that disappeared in a snapshot were actually active but not reachable by our crawler, for instance due to all paths to those services being temporarily unavailable.

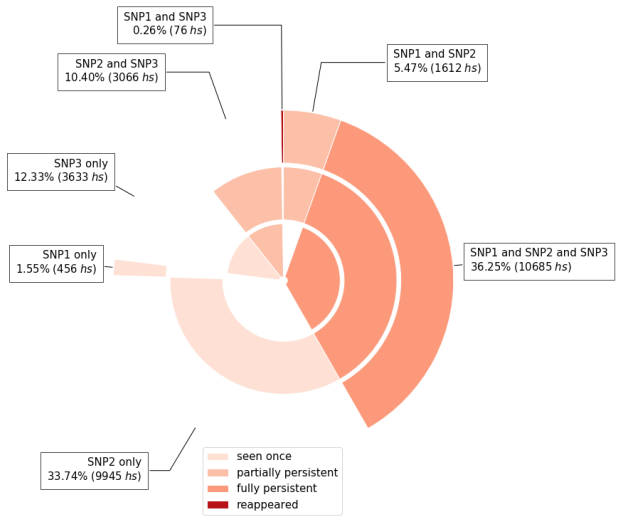


Figure 2: Services persistance over time: inner disc is SNP1, middle disc is SNP2, outer disc is SNP3.

The question now arises of whether hidden services found in two or more snapshots induce the same subgraph in the corresponding snapshots. The answer, summarized in Table 3 by looking at edge density, is no. As a consequence, if we aim at identifying the stable core of our dataset we should not just look at durable hidden services, but we must refer to durable edges. The total amount of durable edges turns out to be 28914, but somewhat surprisingly, one of these edges is isolated from the rest. The common subgraph induced by the set of stable edges consists of two weakly connected components: (i) a giant one, denoted CORE graph in the following, composed of 7669 vertices and 28913 edges, and (ii) a tiny one composed of a single edge connecting the hidden service violet77pvqdmsiy.onion to the hidden service typefacew3ijwkgg.onion.

[9] https://metrics.torproject.org/hidserv-dir-onions-seen.html?start=2017-01-01&end=2017-05-01

Table 3: Density of the subgraphs induced by different nodesets intersections in different graphs.

Nodeset	Density in SNP1	Density in SNP2	Density in SNP3
SNP1 ∩ SNP2	0.000474	0.000461	nd
SNP1 ∩ SNP3	0.000588	nd	0.000507
SNP2 ∩ SNP3	nd	0.000516	0.000435
SNP1 ∩ SNP2 ∩ SNP3	0.000595	0.000584	0.000511

4.2 Global Metrics

The global properties of our four graphs (the three snapshots plus the CORE graph) are summarized in Table 4 and in Table 5. As already mentioned, there is a significant variance in the sizes N and M of the three snapshots, which is however consistent with publicly available aggregated statistics. A common aspect is the lack of a very large hub gathering most of the connections, opposed to the presence of a single vertex that links to, respectively, 44%, 51% and 61% of the whole network. By manual inspection, we found that these onion urls are Tor link directories, not surprisingly. We also checked that the hidden services with greater in- and/or out-degree are generally persistent over the three snapshots, although their rank in the top-degree chart may change. Yet, stability is not a common property of *all* high-degree hidden services, in fact: (i) SNP3 is the only graph including a vertex with in-degree 1464 (which explains the larger C_{in} of that graph), (ii) in the CORE graph the ratios Δ_{in}/N and Δ_{out}/N are comparably smaller with respect to the snapshots, and (iii) in general all parameters strictly related to the presence of hidden services with large in- and/or out-degree (D, Δ_{in}, Δ_{out}, $\langle \deg \rangle$, ρ, C_{in} and C_{out}) appear to be variable over time. The vertex with in-degree 1464 is the hidden service *dhosting4okcs22v.onion* that is a hosting service named `Daniel's Hosting`. Tor users can get a hosting account on the server of the hidden service. The website specifies few rules regarding the contents that can be hosted for the purpose of avoiding illegal or offensive material. The same hidden service contains several sections including a link directory, but each section is registered with a different onion address, *i.e.*, a different hidden service.

Transitivity and clustering coefficients are also variable across time, but in this case the corresponding values for the CORE graph are close to the average of the three snapshots, suggesting that the variance is due to statistical fluctuations in the composition of a network with many volatile nodes. In general the overall frequency of triangles C is substantially in line with the average degree $\langle \deg \rangle$ (as we will better discuss in Section 5), with cycles (T_1) being more frequent than other types of triangles (T_2). The diameter d is stable and logarithmic in N for all graphs, and the ratio of r and d suggests a certain level of symmetry in the graph. The average shortest path length $\langle dist \rangle$ is instead consistently smaller than $\log(N)$, an important factor in determining to what extent the graph resembles a random graph (again, more details will be given in Section 5). Finally, the global efficiencies E_{in} and E_{out}, which should be comparable to $1/\langle dist \rangle$ in uniformly connected networks, are instead diluted by the fact that many pairs of nodes are disconnected (*i.e.*, have no paths connecting each other).

Table 4: Global metrics computed for the services graph of each snapshot

metrics	Graph SNP1	Graph SNP2	Graph SNP3	Graph CORE
$\|V\|$	12829	25308	17460	7669
$\|E\|$	72556	113014	103402	28913
D	0.00044	0.00018	0.00034	0.00049
δ_{in}	1	1	1	0
Δ_{in}	204	262	1464	55
δ_{out}	0	0	0	0
Δ_{out}	5603	12852	10664	2670
$\langle \deg \rangle$	5.65562	4.46554	5.92222	3.77011
ρ	-0.319	-0.32655	-0.16206	-0.37393
C_{in}	0.01546	0.01018	0.08352	0.00668
C_{out}	0.43637	0.50769	0.6105	0.34775
C	0.00943	0.00407	0.00492	0.00873
T_1	0.00617	0.00779	0.00253	0.00535
T_2	0.0039	0.0016	0.00197	0.00356
d	10	12	10	10
r_{in}	5	7	5	0
r_{out}	0	0	0	0
$\langle dist \rangle$	3.79316	4.96028	3.66455	3.98291
E_{in}	0.00549	0.00386	0.02095	0.00371
E_{out}	0.00531	0.00366	0.01965	0.00318

Table 5: Snapshot Data Details

snp	#scc	LSCC size	out-degree 0	in-degree 1
SNP1	12305	466	90.77%	17.478%
SNP2	24433	820	94.74%	43.15%
SNP3	15029	2371	83.32%	24.43%
CORE	7477	169	95.5%	25.09%

4.3 Degree and Centralities

To deepen our understanding of the structure of the Tor Web graph, we now analyze the distribution of a few metrics that quantify the importance of single vertices in the topology of the network and that can be used to gain an insight into the dynamics and the information flow in the graph. Many measures have been introduced in the last 50 years to understand who occupies critical, or *central*, positions in a network [11]. We chose to focus on the in- and out-degree, the PageRank and the betweenness centrality, which are among the most commonly used to describe real-world graphs because they respond to *semantically* different notions of "vertex centrality". The definition of PageRank and betweenness centrality is reported in Table 2.

Figures 3a and 3b show the distributions of, respectively, in-degree and out-degree for all four graphs on a log-log scale. Based on Figure 3a, the in-degree distribution seems to follow a power law decay at least for degrees lying in some intermediate range between ~ 10 and ~ 60. This is not surprising: power law degree distributions are typical in social and web networks, as we will better discuss

in Section 5. To confirm this intuition, we fitted a power law to the distribution using the statistical methods developed in [16]. In particular, we relied on the implementation provided by the POWERLAW python package [3]. POWERLAW autonomously finds a lower-bound $k_{\min}$ for degrees to be fitted, and tries to fit the whole tail unless a context-driven upper-bound $k_{\max}$ is explicitly provided by the user. We also used *Fibonacci binning* [39], as done in a previous work about the surface WWW [29], to show how the distribution looks like if a logarithmic binning is used to smoothen the tail. Two aspects of the obtained fit must be underlined: (i) for SNP2 and SNP3 the α exponent is greater than the threshold 3 that is known to control the variance of the distribution, whereas for SNP1 α is close to 2.9 and for the CORE graph it is 2.7; (ii) the $k_{\min}$ returned by POWERLAW is, respectively, 11, 16, 17 and 5, but the vertices with in-degree greater than or equal to this $k_{\min}$ are only 1624 ($\sim$ 12.67%) in SNP1, 1155 ($\sim$ 4.56%) in SNP2 and 964 ($\sim$ 5.52%) in SNP3, whereas they are 1779 ($\sim$ 23.20%) in the CORE graph, which also has a shorter tail. Summing up, the in-degree provides a further element in support of the intuition that the structure of the CORE graph differs significantly from the snapshots.

Figure 3b makes apparent that the out-degree distribution does not follow a power law. Yet, there are at least two remarkable aspects in this distribution. On the one hand, vertices of out-degree 0 weight 83% up to 95%, according to the graph (this number is reported only in the legend since $k = 0$ is cut-off by the log scale). In other words, a vast majority of Tor's hidden services do not link to any other hidden service. On the other hand, the distribution has a long tail, meaning that the large value of Δ_{out} observed in Section 4.2 is not an isolated case, but rather an evidence of a general trend. To better understand how easily the whole graph can be explored from just a few starting points, in Figure 4 we plot the cumulative percentage of the network that is at distance one from the top out-degree vertices. We see that the top-3 and top-6 out-degree hidden services suffice to reach more than 90% of the graph in just one click in SNP3 and SNP2 respectively. In SNP1 we need the top-20 out-degree hidden services to reach the same percentage of the graph, whereas with the top-6 out-degree services we reach 80% of the nodes. This phenomenon is less evident for the CORE graph, albeit 10 hidden services still contain direct links to more than 80% of the network.

Figures 3c and 3d show the distributions of, respectively, the PageRank and the betweenness centrality for all four graphs on a log-log scale. We opted for a log-log scale in order to make these distributions directy comparable with the in- and out-degree, other than with one another. This is especially important for the PageRank because it has been shown that in many real-world networks (*e.g.*, in scale-free networks) the PageRank distribution "mimics" the in-degree distribution, following a power law with very similar exponent [29]. According to the data points plotted in Figure 3c, this may not seem to be the case for the Tor Web: although it has a heavy tail the decay looks much faster than a power law. However, since PageRank is a continuous metrics, a power law decay can only be appreciated graphically when using a suitable binning. We therefore proceeded exactly as for the in-degree by using POWERLAW to fit the distribution and by applying Fibonacci binning to have a more reliable visual perspective. Unfortunately, the apparently good fit plotted in Figure 3c only regards a minimal portion of the graph: it is only valid for 723 vertices ($\sim$ 5.64%) of SNP1, 772 vertices ($\sim$ 3.05%) of SNP2, 298 vertices ($\sim$ 1.71%) of SNP3 and just 113 vertices ($\sim$ 1.47%) of the CORE graph. For what concerns the betweenness centrality, in all four graphs the long tail and the fast decay are accompanied by more than 90% of the vertices having $B(V) = 0$. This is not surprising since, by definition, all vertices having out-degree 0 must have betweennes 0. Figure 3d partially resembles Figure 3b, suggesting that due to the huge percentage of sinks and to the greater imbalance of the out-degree with respect to the in-degree, in the Tor Web the out-degree impacts on the betweenness of a hidden service more that its in-degree. To confirm or deny this intuition, and more generally to assess the level of correlation between different centrality measures, in Figure 5 we plot the pairwise comparison of (normalized) in-degree, out-degree, PageRank and betweenness centrality.

4.4 Bow-Tie Structure

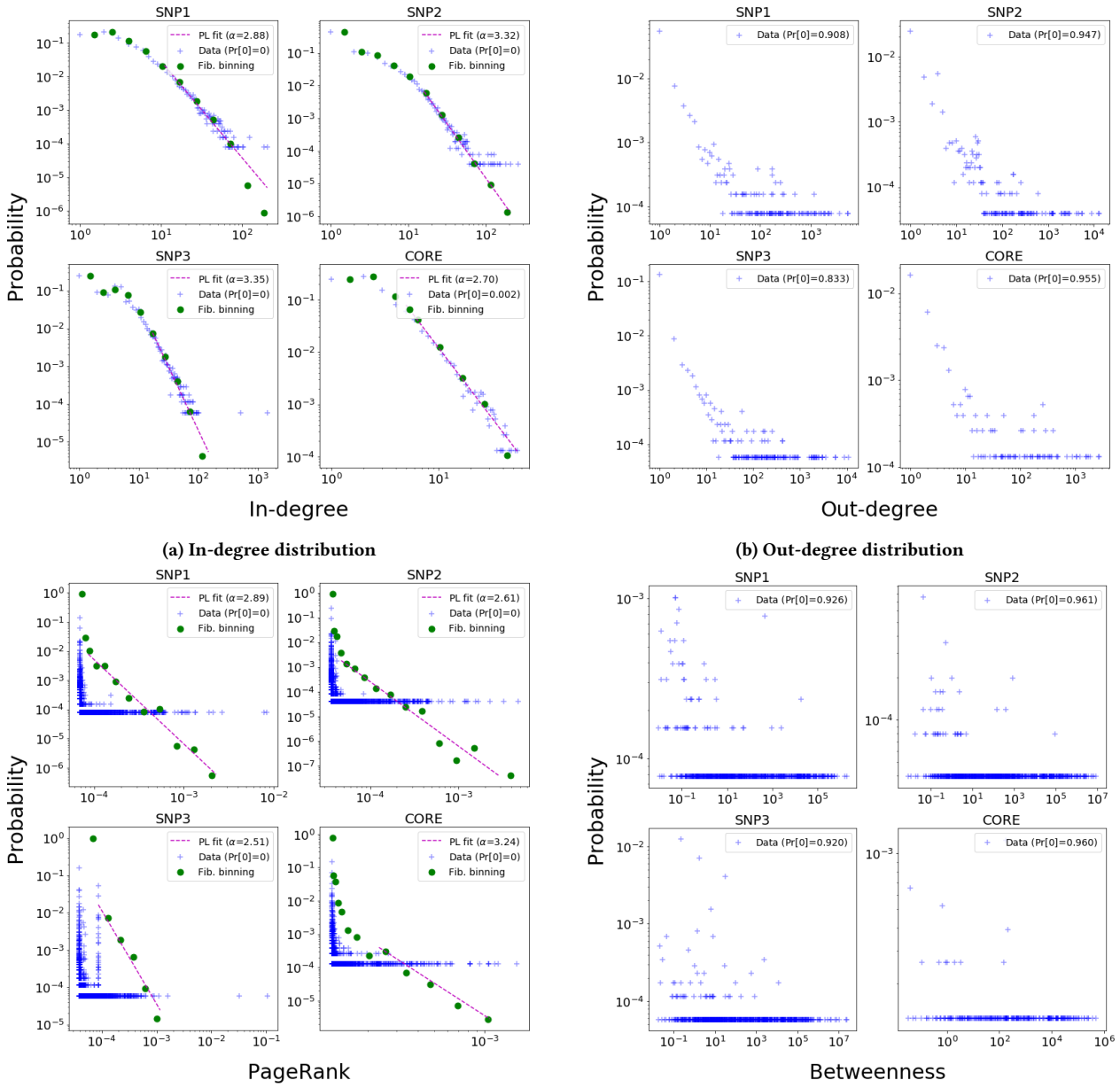
As commonly done to describe Web graphs [12], in Table 6 we provide a bow-tie decomposition of our graphs compared with previous results from the literature. Our findings broadly confirm what emerged in [23], *i.e.*, that the difference between the Tor Web and the WWW is huge and well synthetized by two facts: the LSSC is very small and it lies "on top" of everything else. However, with respect to [23] we implemented a more thorough data collection that brings to the light three novel features of the Tor Web. First, there exists a nonempty set of active hidden services that are completely disconnected from the rest of the graph. Second, the share of the LSCC in the total size of the graph is significantly variable over time, to the point that in SNP3 it is $\sim 4\times$ larger than in the other two snapshots. Finally, the structure of the CORE graph has a few peculiarities: the IN component is non-empty, but instead it is composed of a tiny set of 9 hidden services; the LSCC is even smaller than in the snapshots; the DISCONNECTED component is significanly larger, meaning that in general hyperlinks are more volatile than hidden services.

4.5 Top Hidden Services by Centralities

Considering the pairwise comparison of in-degree, out-degree, PageRank and betweenness of each hidden service shown in Figure 5, we observed that the most interesting services are usually link directories. In SNP1 we find *fhostingesps6bly,*, the service with the top in-degree and PageRank values, that contains an URL redirection[10] to the Hidden Wiki whose current onion address is *zqktlwi4i34kbat3*. The Hidden Wiki is a Tor link directory, probably the most famous. In SNP1 we find also *underdj5ziov3ic7* that is the service with the top betweenness and out-degree values. This service contains a Tor link directory named UnderDir - The Undernet Directory. The last service in SNP1 is *blockchainbdgpzk* that is the service with the second PageRank and in-degree values. This service contains an URL redirection to a surface website of a company named Blockchain Luxembourg S.A.R.L. that offers services related to digital currencies.

In SNP2 we find *tt3j2x4k5ycaa5zt* that is the service with the top PageRank and second in-degree values; it contains a personal

[10]The HTTP Status Code used for an URL redirection/URL forwarding is a 3XX status code



(a) In-degree distribution

(b) Out-degree distribution

(c) PageRank distribution

(d) Betweenness distribution

Figure 3: Probability distribution of centrality measures.

website named Daniel's Home. In the website a section is dedicated to a collection of Tor links. In SNP2 we find also *zlal32teyptf4tvi* that is the top betweenness and second out-degree service. This service is a Tor link directory named Fresh Onions. The last service in SNP2 is *vj5wxqmjaes2bae5* that has the top out-degree value, it contains a Tor link directory.

In SNP3 we find *dhosting4okcs22v* that is the service with top PageRank, betweenness and in-degree values; it contains a hosting service named Daniel's Hosting. This is a section of the website we found in SNP2 named Daniel's Home. It uses a different onion address, but it is actually the same website. The last service in SNP2 is *zlal32teyptf4tvi* that is the service with the top out-degree value containing the Tor link directory named Fresh Onions, that we found in SNP2.

Finally in CORE we find *blockchainbdgpzk* and *underdj5ziov3ic7* that are respectively the services with the top PageRank, second in-degree service and the top betweenness, out-degree values. We already discussed these two onion addresses for SNP1. In CORE we find also *grams7enufi7jmdl* that is the service with the second in-degree value. This service contains a Tor search engine focused on

Table 6: Bow-Tie structure

Component	WWW from [29]		Tor from [23]		SNP1		SNP2		SNP3		CORE	
	# nodes	%	# nodes	%	# nodes	%	# nodes	%	# nodes	%	# nodes	%
LSCC	22.3M	51.94%	297	4.14%	466	3.63%	820	3.24%	2371	13.58%	169	2.2%
IN	3.3M	7.65%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	9	0.12%
OUT	13.3M	30.98%	6881	95.86%	12312	95.94%	24468	96.68%	15057	86.24%	7353	95.88%
TUBES	17K	0.04%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
TENDRILS	514k	1.2%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
DISCONNECTED	3.5M	8.2%	0	0.0%	55	0.43%	20	0.08%	32	0.18%	138	1.8%

LSCC is the largest strongly connected component.
IN is the set of nodes $v \in V \setminus$ LSCC such that there is a path from v to LSCC.
OUT is the set of nodes $v \in V \setminus$ LSCC such that there is a path from LSCC to v.
TUBES is the set of nodes $v \in V \setminus$ (LSCC $\cup$ IN $\cup$ OUT) such that there is a path from IN to v as well as a path from v to OUT.
TENDRILS is the set of nodes $v \in V \setminus$ (LSCC $\cup$ IN $\cup$ OUT) such that there is either a path from IN to v or a path from v to OUT, but not both.
DISCONNECTED is the set of all other nodes $v \in V \setminus$ (LSCC $\cup$ IN $\cup$ OUT $\cup$ TUBES $\cup$ TENDRILS).

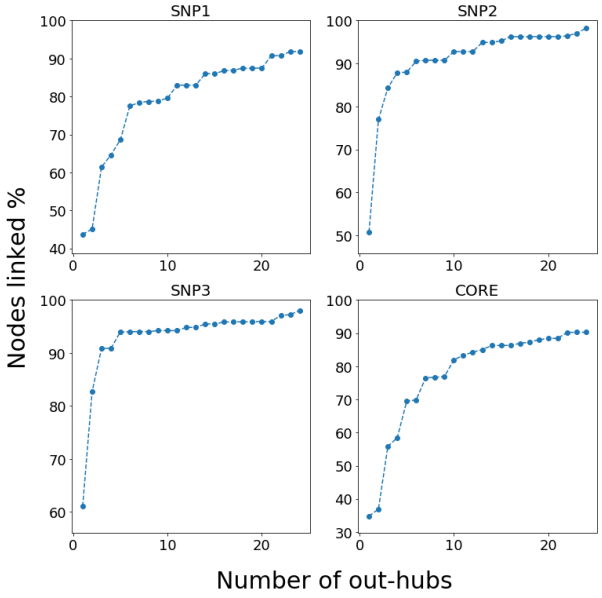


Figure 4: Cumulative percentage of the graph linked by the top out-degree vertices.

Tor marketplaces. The last service in CORE is *zqktlwi4fecvo6ri* that is the service with the second betweenness value and it contains an URL redirection to the Hidden Wiki.

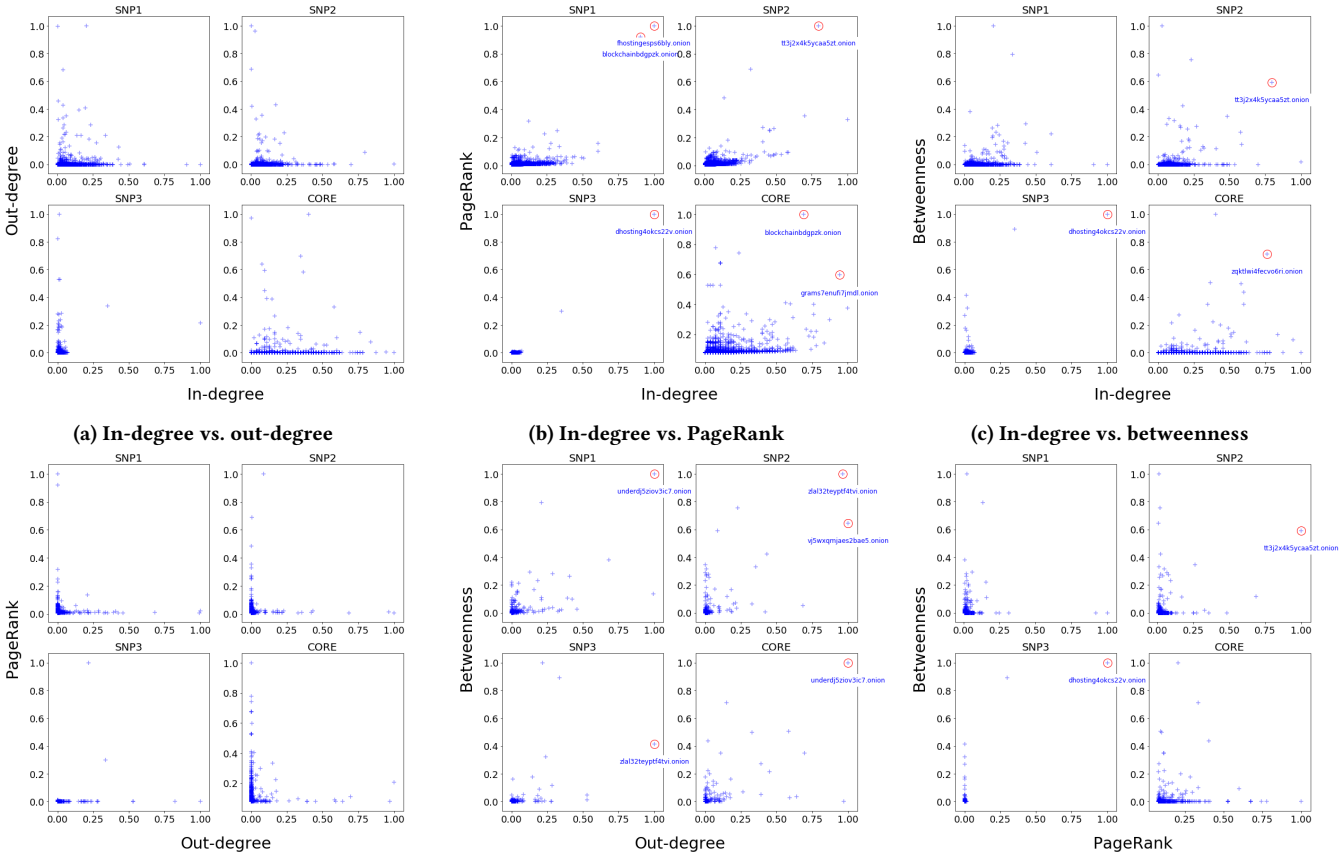
5 DISCUSSION

In this Section we read and discuss our findings in the light of the body of work on real world complex networks. Aiming at assessing to which extent the Tor Web graph fits the three most-known generative models for random graphs – Erdos-Renyi (ER) [20], Watts-Strogatz (WS) [40], Barabasi-Albert (BA) [5] – we focus on three properties that are especially informative: the average shortest path length (or average distance), the clustering coefficient (or transitivity) and the shape of the degree distribution. Notably, these three characteristics are also known to be discriminatory in many practical settings, such as: for predicting the growth dynamics of a network [2], for controlling the spreading of viruses/rumors [32], or for determining the robustness against random node failures [13].

For what concerns the average distance, the question is whether the Tor graph looks like a *small world* or even an *ultra-small world* network. In small world networks the average distance satisfies $\langle dist \rangle \propto \ln(N)$, whereas in ultra-small networks $\langle dist \rangle \propto \ln\ln(N)$. Although the latter are asymptotic estimates, our numbers suggest that Tor belongs, at least, to the class of small world networks: for all four graphs considered (the three snapshots and the CORE) $\langle dist \rangle$ satisfies $3.5 < \langle dist \rangle < 5$, whereas $2 < \ln\ln(N) < 2.5$ and $8.5 < \ln(N) < 10.5$.

The global clustering coefficient measures the frequency of closed triangles in the graph, thus representing an indicator of the existence of some level of correlation in the adjacency patterns of neighboring vertices. If edges occur independently and uniformly at random, the clustering coefficient C satisfies $C \propto \frac{\langle \deg \rangle}{N}$, where $\langle \deg \rangle$ is the average degree of the network. Again, although the latter is only an asymptotic estimate, our findings speak in favor of the existence of a positive correlation for Tor edges. For directed graphs, other than the global (undirected) clustering coefficient it is possible to consider a few types of directed transitivities: we chose two such metrics, denoted T_1 and T_2 and defined in Table 2. Similarly, the average degree can be computed both ignoring or considering edge directions, and the value reported in Table 4 for $\langle \deg \rangle$ is the average number of in-bound or (equivalently) out-bound edges, thus C should be compared with $\frac{2\langle \deg \rangle}{N}$ in our case. In all four Tor graphs, C is one order of magnitude greater than $\frac{2\langle \deg \rangle}{N}$, and both T_1 and T_2 are one order of magnitude greater than $\frac{\langle \deg \rangle}{N}$.

Finally, when looking at the degree distribution of a network, the first aspect to consider is whether it has a heavy tail, which is symptomatic of the tendency of nodes to connect to "authoritative" hubs. By looking at Figures 3a and 3b it is clear that this is the case for the Tor Web. However, while the in-degree distribution seems to follow a power law for all four graphs, albeit with different



(a) In-degree vs. out-degree **(b) In-degree vs. PageRank** **(c) In-degree vs. betweenness**

(d) Out-degree vs. PageRank **(e) Out-degree vs. betweenness** **(f) PageRank vs. betweenness**

Figure 5: Pairwise comparison between centrality measures.

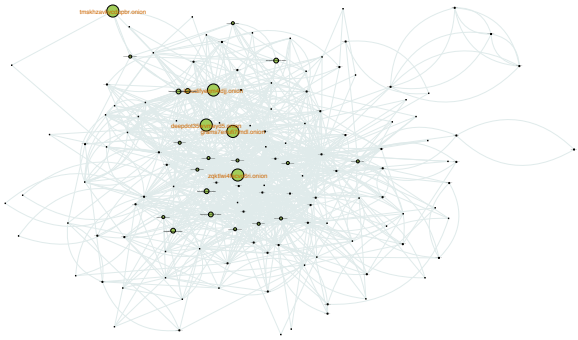
exponents, the out-degree shows an even slower decay and an even longer tail.

Our findings make it clear that the Tor Web is not an ER graph. In the ER random graph model with N vertices and with expected degree $\langle \deg \rangle$, the expected average distance satisfies $\langle dist \rangle \propto \ln(N)$, but the expected global clustering coefficient is $C \propto \frac{\langle \deg \rangle}{N}$ and the degree distribution has no heavy tail. The WS model predicts that $\langle dist \rangle \propto \ln(N)$ and that $C \gg \frac{\langle \deg \rangle}{N}$, thus suiting our findings quite well. Yet, the values of C measured in our four graphs are not as large as in other real-world small world networks [5, 32, 40]. Additionally, the WS model alone does not predict the observed power law distribution of the in-degree. The power law degree distribution is the defining property of the BA model. However, the BA model theoretically predicts the existence of two regimes according to whether the power law exponent α satisfies $2 < \alpha < 3$ or $\alpha > 3$. In the former case, the variance of the distribution diverges and the network is ultra-small. In the latter, the variance of the distribution is finite and the network is small. The existence of this threshold is not visible in our findings: SNP1 and CORE have $2 < \alpha < 3$, whereas SNP2 and SNP3 have $\alpha > 3$, but all four graphs have very similar average distance and degree distribution. Although real networks cannot have an infinite variance, if N is large enough, node degrees should span several orders of magnitude if and only if $2 < \alpha < 3$, but this seems not be the case for the out-degree in Tor.

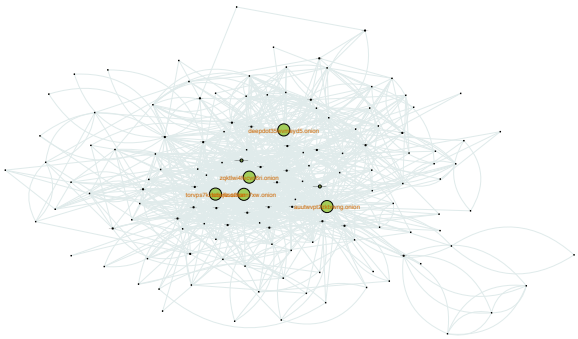
Summing up, Tor shows a blend of different features and a deeper analysis is needed to understand whether any existing generative model matches its characteristics. Among Tor's unique features, we observe that the overall structure is clearly dominated by the presence of a small number of *out-hubs*, which connect both central and peripheral nodes. It is also worth noting that the sizes of the LSCCs are relatively small compared to those reported by other studies on the surface WWW or social networks [5, 27, 32]. It is interesting that the nodes in the LSCC core are persistent in the three snapshots, suggesting that some subnet inside Tor could have specific properties and even a different structure with respect to the whole network.

Many questions that are still open require further analysis. Tor is built with anonymity in mind thus many hidden services are supposedly not interested in having visibility, yet some of its topological features, *e.g.* the radius and diameter, suggest the existence of a peculiar mechanism that leads to the growth of the network. On the practical side, one may be especially interested in understanding how the existence of large *out-hubs* impacts on the topological

properties of the network, and whether topologically similar nodes host analogous contents. To provide a few insights in this regards, we removed from the CORE set the 10 services with maximum out-degree and analyzed the LSCC of the obtained graph, depicted in Figure 6. The most remarkable finding is that, albeit the average shortest path remains almost unchanged, the obtained network has a clustering coefficient more than 20 times larger than that of a random graph of equal size. We also found that *out-hubs* are almost all *hidden directories* and we have some evidence that other centrality measures can be related to the content of the hidden services. In particular, as reported in Table 7, 4 out of the top 5 PageRank services are related to marketplaces, whereas high betweenness seems not equally characteristic of a specific category of services.



(a) Node size proportional to their PageRank.



(b) Node size proportional to their betweenness centrality.

Figure 6: The LSCC of the CORE graph.

6 CONCLUSIONS

This paper studied three sets of crawling data collected over three different time frames as well as their common "stable" core. It provided a deep characterization of the topology of the Tor services graph, identifying structural and temporal features and further assessing the persistence of hidden services and hyperlinks.

While previous work only focused on the volatility of Tor hidden services, thanks to a graph-oriented perspective we were also able to assess the persistence of Tor hyperlinks. This led to the key finding that edges are more volatile than nodes in the Tor Web graph, as proved by the fact that hidden services shared by two or more snapshots do not induce the same subgraph in these snapshots. Additionally, we observed that the LSCC of the CORE graph is persistent in all three snapshots. Compared with the generally high volatility of the Tor network, this is the first evidence that Tor may be comprised of different layers, each with a precise role in the connectivity patterns of the network, and possibly with different inter- and intra-layer structures.

onion	PAGERANK	Topic
grams7enufi7jmdl	0.053	Market SE
deepdot35wvmeyd5	0.048	Market
lchudifyeqm4ldjj	0.046	Market
zqktlwi4fecvo6ri	0.045	Hidden Wiki
tmskhzavkycdupbr	0.041	Market
onion	BETWEENNESS	Topic
auutwvpt2zktxwng	6331	Directory
zqktlwi4fecvo6ri	6105	Hidden Wiki
deepdot35wvmeyd5	3036	Market
torpress2sarn7xw	2766	News/Blogs
torvps7kzis5ujfz	2396	News/Blogs

Table 7: Best 5 nodes with respect to PageRank and betweenness centrality.

We computed several topological metrics on the Tor snapshots and compared them to well-known network models (ER, WS, BA). None of those models appears to be suitable to accurately represent Tor. A small number of *out-hubs* connecting both central and peripheral nodes dominates the structure of the graph, whereas LSCCs are (relatively/due proportions made) smaller than those of the surface Web or social networks. By removing the *out-hubs* with higher degree, the clustering coefficient of the network grows but the average shortest path remains almost constant: this again suggests that additional insights into Tor's dynamics could be obtained by removing and/or isolating specific subnets. Centrality metrics also indicate that there could be some interesting relation among node role/position and content: nodes with higher degree are almost always link directories, whereas we found that four out of five nodes in the top PageRank are related to marketplaces. While these results do not suffice to draw a clear picture, they surely indicate that further research must be carried out [11].

We believe the results presented here will foster a larger discussion on the topic, and will be a useful reference for evaluation and comparison against other real-world graphs.

REFERENCES

[1] Dimitris Achlioptas, Aaron Clauset, David Kempe, and Cristopher Moore. 2009. On the Bias of Traceroute Sampling: Or, Power-law Degree Distributions in Regular Graphs. *J. ACM* 56, 4, Article 21 (July 2009), 28 pages. https://doi.org/10.1145/1538902.1538905

[2] Réka Albert, Hawoong Jeong, and Albert-László Barabási. 1999. Internet: Diameter of the world-wide web. *nature* 401, 6749 (1999), 130.

[3] Jeff Alstott, Ed Bullmore, and Dietmar Plenz. 2014. Powerlaw: a Python package for analysis of heavy-tailed distributions. *PloS one* 9, 1 (2014), e85777.

[11] Further information is available here http://www.cranic.it/tor.html, here 757qx3mfy4opg73o.onion, and here dym3ubprp66kbj33o22q6mehjwvixknrblndb65oi5sk2v2fpx4enrad.onion

[4] Robert Annessi and Martin Schmiedecker. 2016. NavigaTor: Finding Faster Paths to Anonymity. In *IEEE European Symposium on Security and Privacy (Euro S&P)*. IEEE.

[5] Albert-László Barabási and Réka Albert. 1999. Emergence of Scaling in Random Networks. *Science* 286, 5439 (1999), 509–512. https://doi.org/10.1126/science.286.5439.509

[6] Massimo Bernaschi, Alessandro Celestini, Stefano Guarino, and Flavio Lombardi. 2017. Exploring and Analyzing the Tor Hidden Services Graph. *ACM Trans. Web* 11, 4, Article 24 (jul 2017), 24:1–24:26 pages. https://doi.org/10.1145/3008662

[7] Alex Biryukov, Ivan Pustogarov, Fabrice Thill, and Ralf-Philipp Weinmann. 2014. Content and Popularity Analysis of Tor Hidden Services. In *Distributed Computing Systems Workshops (ICDCSW), 2014 IEEE 34th International Conference on*. 188–193. https://doi.org/10.1109/ICDCSW.2014.20

[8] Alex Biryukov, Ivan Pustogarov, and Ralf-Philipp Weinmann. 2013. Trawling for Tor Hidden Services: Detection, Measurement, Deanonymization. In *Proceedings of the 2013 IEEE Symposium on Security and Privacy (SP '13)*. IEEE Computer Society, Washington, DC, USA, 80–94. https://doi.org/10.1109/SP.2013.15

[9] Paolo Boldi, Andrea Marino, Massimo Santini, and Sebastiano Vigna. 2014. BUbiNG: Massive crawling for the masses. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*. 227–228.

[10] Anthony Bonato. 2005. A Survey of Models of the Web Graph. In *Combinatorial and Algorithmic Aspects of Networking*, Alejandro Lopez-Ortiz and Angelem Hamel (Eds.). Lecture Notes in Computer Science, Vol. 3405. Springer Berlin Heidelberg, 159–172. https://doi.org/10.1007/11527954_16

[11] Stephen P. Borgatti and Martin G. Everett. 2006. A Graph-theoretic perspective on centrality. *Social Networks* 28, 4 (2006), 466 – 484. https://doi.org/10.1016/j.socnet.2005.11.005

[12] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. 2000. Graph structure in the Web. *Computer Networks* 33, 1-6 (2000), 309 – 320. https://doi.org/10.1016/S1389-1286(00)00083-9

[13] Duncan S. Callaway, M. E. J. Newman, Steven H. Strogatz, and Duncan J. Watts. 2000. Network Robustness and Fragility: Percolation on Random Graphs. *Phys. Rev. Lett.* 85 (Dec 2000), 5468–5471. Issue 25. https://doi.org/10.1103/PhysRevLett.85.5468

[14] Alessandro Celestini and Stefano Guarino. 2017. Design, Implementation and Test of a Flexible Tor-oriented Web Mining Toolkit. In *Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics (WIMS '17)*. ACM, New York, NY, USA, Article 19, 10 pages. https://doi.org/10.1145/3102254.3102266

[15] Nicolas Christin. 2013. Traveling the Silk Road: A Measurement Analysis of a Large Anonymous Online Marketplace. In *Proceedings of the 22Nd International Conference on World Wide Web (WWW '13)*. ACM, New York, NY, USA, 213–224. https://doi.org/10.1145/2488388.2488408

[16] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. 2009. Power-law distributions in empirical data. *SIAM review* 51, 4 (2009), 661–703.

[17] Manlio De Domenico and Alex Arenas. 2017. Modeling structure and resilience of the dark network. *Phys. Rev. E* 95 (Feb 2017), 022313. Issue 2. https://doi.org/10.1103/PhysRevE.95.022313

[18] Debora Donato, Stefano Leonardi, Stefano Millozzi, and Panayiotis Tsaparas. 2008. Mining the inner structure of the Web graph. *Journal of Physics A: Mathematical and Theoretical* 41, 22 (2008), 224017. http://stacks.iop.org/1751-8121/41/i=22/a=224017

[19] Paul Erdős and Alfréd Rényi. 1959. On random graphs. *Publicationes Mathematicae Debrecen* 6 (1959), 290–297.

[20] P. ErdÅŠs and A RÃľnyi. 1960. On the Evolution of Random Graphs. In *PUBLICATION OF THE MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES*. 17–61.

[21] Massimo Franceschet. 2011. PageRank: Standing on the Shoulders of Giants. *Commun. ACM* 54, 6 (June 2011), 92–101. https://doi.org/10.1145/1953122.1953146

[22] Shalini Ghosh, Ariyam Das, Phil Porras, Vinod Yegneswaran, and Ashish Gehani. 2017. Automated Categorization of Onion Sites for Analyzing the Darkweb Ecosystem. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. ACM, New York, NY, USA, 1793–1802. https://doi.org/10.1145/3097983.3098193

[23] Virgil Griffith, Yang Xu, and Carlo Ratti. 2017. Graph Theoretic Properties of the Darkweb. *arXiv preprint arXiv:1704.07525* (2017).

[24] Rob Jansen, Kevin Bauer, Nicholas Hopper, and Roger Dingledine. 2012. Methodically Modeling the Tor Network. In *Proceedings of the 5th USENIX Conference on Cyber Security Experimentation and Test (CSET'12)*. USENIX Association, Berkeley, CA, USA, 8–8. http://dl.acm.org/citation.cfm?id=2372336.2372347

[25] Rohit Khare, Doug Cutting, Kragen Sitaker, and Adam Rifkin. 2004. Nutch: A flexible and scalable open-source web search engine. *Oregon State University* 1 (2004), 32–32.

[26] JonM. Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and AndrewS. Tomkins. 1999. The Web as a Graph: Measurements, Models, and Methods. In *Computing and Combinatorics*, Takano Asano, Hideki Imai, D.T. Lee, Shin-ichi Nakano, and Takeshi Tokuyama (Eds.). Lecture Notes in Computer Science, Vol. 1627. Springer Berlin Heidelberg, 1–17. https://doi.org/10.1007/3-540-48686-0_1

[27] Ravi Kumar, Jasmine Novak, and Andrew Tomkins. 2010. *Structure and Evolution of Online Social Networks*. Springer New York, New York, NY, 337–357. https://doi.org/10.1007/978-1-4419-6515-8_13

[28] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, D Sivakumar, Andrew Tomkins, and Eli Upfal. 2000. Stochastic models for the Web graph. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*. 57–65. https://doi.org/10.1109/SFCS.2000.892065

[29] Oliver Lehmberg, Robert Meusel, and Christian Bizer. 2014. Graph Structure in the Web: Aggregated by Pay-level Domain. In *Proceedings of the 2014 ACM Conference on Web Science (WebSci '14)*. ACM, New York, NY, USA, 119–128. https://doi.org/10.1145/2615569.2615674

[30] Robert Meusel, Sebastiano Vigna, Oliver Lehmberg, and Christian Bizer. 2015. The Graph Structure in the Web âĂŞ Analyzed on Different Aggregation Levels. *The Journal of Web Science* 1, 1 (2015), 33–47. https://doi.org/10.1561/106.00000003

[31] Gordon Mohr, Michael Stack, Igor Ranitovic, Dan Avery, and Michele Kimpton. 2004. An Introduction to Heritrix An open source archival quality web crawler. In *In IWAWâĂŹ04, 4th International Web Archiving Workshop*. Citeseer.

[32] Mark Newman, Albert-Laszlo Barabasi, and Duncan J. Watts. 2006. *The Structure and Dynamics of Networks: (Princeton Studies in Complexity)*. Princeton University Press, Princeton, NJ, USA.

[33] M. E. J. Newman. 2003. Mixing patterns in networks. *Phys. Rev. E* 67, 2 (Feb 2003), 026126. https://doi.org/10.1103/PhysRevE.67.026126

[34] Gareth Owen and Nick Savage. 2015. The Tor dark net. (2015).

[35] Gareth Owen and Nick Savage. 2016. Empirical analysis of Tor hidden services. *IET Information Security* 10, 3 (2016), 113–118.

[36] Iskander Sanchez-Rola, Davide Balzarotti, and Igor Santos. 2017. The Onions Have Eyes: A Comprehensive Structure and Privacy Analysis of Tor Hidden Services. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1251–1260. https://doi.org/10.1145/3038912.3052657

[37] M. Ángeles Serrano, Ana Maguitman, Marián Boguñá, Santo Fortunato, and Alessandro Vespignani. 2007. Decoding the Structure of the WWW: A Comparative Analysis of Web Crawls. *ACM Trans. Web* 1, 2, Article 10 (Aug. 2007). https://doi.org/10.1145/1255438.1255442

[38] Martijn Spitters, Stefan Verbruggen, and Mark van Staalduinen. 2014. Towards a Comprehensive Insight into the Thematic Organization of the Tor Hidden Services. In *Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint*. 220–223. https://doi.org/10.1109/JISIC.2014.40

[39] Sebastiano Vigna. 2013. Fibonacci binning. *arXiv preprint arXiv:1312.3749* (2013).

[40] Duncan J. Watts. 1999. *Small Worlds: The Dynamics of Networks Between Order and Randomness*. Princeton University Press, Princeton, NJ, USA.