World Wide Web (2022) 25:1287–1313
https://doi.org/10.1007/s11280-022-01044-z

Onion under Microscope: An in-depth analysis of the Tor Web

Massimo Bernaschi[1] · Alessandro Celestini[1] · Marco Cianfriglia[1,2] · Stefano Guarino[1] · Flavio Lombardi[1] · Enrico Mastrostefano[1]

Received: 27 March 2020 / Revised: 7 October 2021 / Accepted: 10 March 2022 /
Published online: 1 April 2022

Abstract
Tor is an open source software that allows accessing various kinds of resources, known as hidden services, while guaranteeing sender and receiver anonymity. Tor relies on a free, worldwide, overlay network, managed by volunteers, that works according to the principles of onion routing in which messages are encapsulated in layers of encryption, analogous to layers of an onion. The Tor Web is the set of web resources that exist on the Tor network, and Tor websites are part of the so-called dark web. Recent research works have evaluated Tor security, its evolution over time, and its thematic organization. Nevertheless, limited information is available about the structure of the graph defined by the network of Tor websites, not to be mistaken with the network of nodes that supports the onion routing. The limited number of entry points that can be used to crawl the network, makes the study of this graph far from being simple. In the present paper we analyze two graph representations of the Tor Web and the relationship between contents and structural features, considering three crawling datasets collected over a five-month time frame. Among other findings, we show that Tor consists of a tiny strongly connected component, in which link directories play a central role, and of a multitude of services that can (only) be reached from there. From this viewpoint, the graph appears inefficient. Nevertheless, if we only consider mutual connections, a more efficient subgraph emerges, that is, probably, the backbone of social interactions in Tor.

Keywords Tor · Web graph · Dark web · Complex networks

✉ Alessandro Celestini
a.celestini@iac.cnr.it

Massimo Bernaschi
m.bernaschi@iac.cnr.it

Marco Cianfriglia
mcianfriglia@uniroma3.it

Stefano Guarino
s.guarino@iac.cnr.it

Flavio Lombardi
f.lombardi@iac.cnr.it

Enrico Mastrostefano
e.mastrostefano@iac.cnr.it

[1] Institute for Applied Computing, National Research Council of Italy, Via dei Taurini 19, Rome, Italy

[2] Department of Mathematics and Physics, Roma Tre University, Largo San Leonardo Murialdo 1, Rome, Italy

1 Introduction

"Dark web" is a generic term for the subset of the Web that, other than being non-indexed by popular search engines, is accessible only through specific privacy-preserving browsers and overlay networks. Those networks, often called *darknets*, implement suitable cryptographic protocols to the purpose of keeping anonymous the identity of both the services offering contents and the users enjoying them. The best known and most widespread of them is probably Tor, which takes its name from The Onion Routing protocol it is based upon. Tor guarantees privacy and anonymity by redirecting traffic through a set of *relays*, each adding a layer of encryption to the data packets they forward. The equivalent of a domain on the surface Web is called Hidden Service (HS) in Tor.

Past research on the Tor network has evaluated its security [8], evolution [22], and thematic organization [36]. Nevertheless, an in depth study of Tor's characteristics is difficult due to the limited number of Tor entry points on the surface web. In this paper, building on and extending over previous results on the topic [5, 6], we aim at better characterizing the Tor Web by analyzing three crawling datasets collected over a five-month time frame. In line with previous work on the WWW [11] and with a recent trend for criminal networks and dark/deep web [5, 13, 20, 34], we investigate Tor as a *complex system*, shedding new light on usage patterns as well as dynamics and resilience of the Tor Web. We consider the Tor Web graph aggregated by HS, *i.e.*, the network of Tor HSs connected by hyperlinks – not to be mistaken with the network of Tor relays. We analyze the topology of two different graph representations of the Tor Web – directed and undirected – also using local properties of the graphs to characterize the role that different services play in the network. Relying on a large dataset of manually tagged HSs [2], we relate a few structural properties with the thematic organization of Tor's web content.

Along with the three *snapshot* graphs induced by the three crawling data sets, we also consider an *intersection* graph and an *union* graph, in an effort to discriminate intrinsic features from noise. As a side effect, the present paper also addresses several open questions about the persistence of the Tor Web, showing the actual changes that took place in the quality, quantity and shape of available services and in their interconnections over the considered time span.

Overall, Tor comes out having significant structural differences with respect to the WWW. Our main findings may be summarized as follows:

- The Tor Web is a network which resembles a *small world* one but is somehow *inefficient*, consisting of a tiny strongly connected component (SCC) surrounded by a multitude of services that can be reached from the SCC but do not allow getting back to it.
- The stable core of the Tor Web is mostly composed of in- and out-hubs, whereas the periphery is highly volatile. The in- and out-hubs are generally separate services in Tor.
- The (relatively small) undirected subgraph of the Tor Web, obtained only considering *mutual* connections, is quite efficient despite it lacks most of the features of a small world network. As a matter of fact, the undirected graph better preserves the social organization of the graph, such as its community structure, which appears to be generally stable and, as such, meaningful.
- Both the volatility of Tor's HSs and the tendency of the HSs to cluster together are unrelated to the services' content.
- With a few exceptions, the topological metrics are scarsely informative of the activity occurring on a service; however, the "hubbiness" of a HS may be of some help in detecting "suspicious" activities (as defined in [1]).

To the best of our knowledge, the amount of data we collected for the study of the Tor Web exceeds previous efforts reported in the literature [6, 7, 20, 36], making possible an in-depth analysis.

1.1 Related Work

Interesting works studying the topology of the underlying network and/or semantically analyzing Tor contents have appeared so far.

Biryukov et al. [8] managed to collect a large number of hidden service descriptors by exploiting a presently-fixed Tor vulnerability to find out that most popular hidden services were related to botnets. Owen et al. [32] reported over hidden services persistence, contents, and popularity, by operating 40 relays over a 6 month time frame.

ToRank [1], by Al-Nabki et al. is an approach to rank Tor hidden services. The authors collected a large Tor dataset called DUTA-10K extending the previous Darknet Usage Text Address (DUTA) dataset [2]. The ToRank approach selects nodes relevant to the Tor network robustness. DUTA-10K analysis reveals that only 20% of the accessible hidden services are related to suspicious activities. It also shows how domains related to suspicious activities usually present multiple clones under different addresses. Zabihimayvan et al. [39] evaluate the contents of English Tor pages by performing a topic and network analysis on crawling-collected data composed of 7,782 pages from 1,766 unique onion domains. They classify 9 different domain types according to the information or service they host. Further, they highlight how some types of domains intentionally isolate themselves from the rest of Tor. Contrary to [1], their measurements suggest how marketplaces of illegal drugs and services emerge as the dominant type of Tor domain. Similarly, Takaaki et al. [37] analyzed a large amount of onion domains obtained using the Ichidan search engine and the Fresh Onions site. They classified every encoutered onion domain into 6 categories, creating a directed graph and attempting to determine the relationships and characteristics of each instance. Ghosh et al. [18] employed another automated tool to explore the Tor network and analyze the contents of onion sites for mapping onion site contents to a set of categories, and clustered Tor services to categorize onion content. The main limitation of that work is that it focused on page contents/semantics, and did not consider network topology.

A few research works focus on Tor's illegal marketplaces. Duxbury et al. [16] examine the global and local network structure of an encrypted online drug distribution network. Their aim is to identify vendor characteristics that can help explain variations in the network structure. Their study leverages structural measures and community detection analysis to characterize the network structure. Norbutas et al. [31] made use of publicly available crawls of a single cryptomarket (Abraxas) during 2015 and leveraged descriptive social network analysis and Exponential Random Graph Models (ERGM) to analyze the structure of the trade network. They found out the structure of the online drug trade network to be primarily shaped by geographical boundaries, leading to strong geographic clustering, especially strong between continents and weaker for countries within Europe. As such, they suggest that cryptomarkets might be more localized and less international than thought before. Christin et al. [13] collected crawling data on specific Tor hidden services over an 8 month lifespan. They evaluated the evolution/persistence of such services over time, and performed a study on the contents and the topology of the explored network. The main difference with our 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. [15], used the data collected in [4] to study the topology of the Tor network. They gave a characterization of the topology of this 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. Griffith et al. [20] performed a topological analysis of the Tor hidden services graph. They crawled Tor using the *scrapinghub.com* commercial service through the *tor-2web* proxy onion link. Interestingly, they reported that more than 87% of dark websites never link to another site. The main difference with our 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 network analysis (we evaluate a far larger set of network characteristics).

So far, one of the largest Tor dataset collected from an automated Tor network exploration is due to Bernaschi et al. [5]. They aimed at relating semantic contents similarity with Tor topology, searching for smaller connected components that exhibit a larger semantic uniformity. Their results show that the Tor Web is very topic-oriented, with most pages focusing on a specific topic, and only a few pages dealing with several different topics. Further work [6] by the same authors features a very detailed network topology study investigating similarities and differences from surface Web and applying a novel set of measures to the data collected by automated exploration. They show that no simple graph model fully explains Tor's structure and that out-hubs govern the Tor's Web structure.

1.2 Roadmap

The rest of the paper is organized as follows. In Section 2 we describe: (i) our dataset, including statistics about the organization of the hidden services as websites (tree map, amount of characters and links); (ii) the DUTA dataset we used for content analysis. In Section 3 we describe how we extracted our graph representations from the available data and we recall the definition of all graph-related notation and metrics used throughout the paper. In Section 4 we discuss and present the results of our in-depth analysis of the Tor Web, carried out through a set of structural measures and statistics. We study properties such as bow-tie decomposition, global and local (i.e., vertex-level) metrics, degree distributions, community structure, and content related distribution and metrics. Finally, we draw conclusions in Section 5.

2 Data

The present paper analyzes a dataset that is the result of three independent six-week runs of our customized crawler, resulting in three "snapshots" of the Tor Web: SNP1, SNP2 and SNP3. The design of the crawler and the outcome of the scraping procedures are reported in Appendix 1 and more extensively discussed in [6, 12].

It is quite common to analyze a dataset obtained by crawling the web. Yet, it must be kept in mind that the analysis may be susceptible to fluctuations due to the order in which pages have been first visited – and, hence, not revisited thereafter [26]. In the case of the Tor Web, the issue is exacerbated by the renowned volatility of Tor hidden services [7, 8, 32]. By executing three independent scraping attempts over five months,

we aimed at making our analysis more robust and at telling apart "stable" and "temporary" features of the Tor Web.

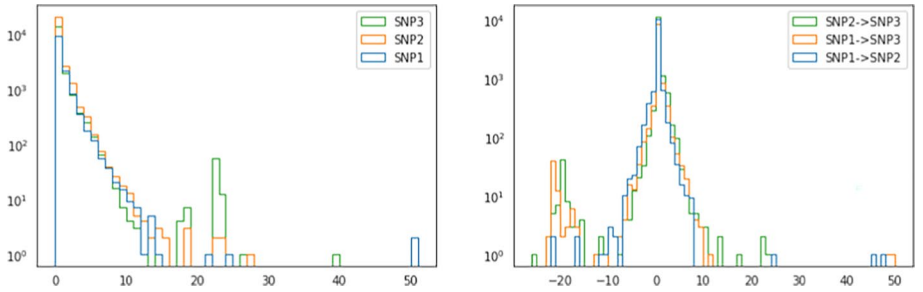
In total, we reached millions of onion pages (more than 3 millions in the second run alone) and almost 30 thousands distinct hidden services. The distribution of these hidden services across the three snapshots is reported in Table 1. Albeit active services may temporarily appear offline to the crawler (*e.g.*, due to all paths to those services being unavailable), these statistics are quite informative about the volatility of the Tor web. Just 10685 onion URLs were successfully reached by all three crawling runs. It is quite likely that those hidden services were durably present over the considered five months time frame; they account for, respectively, 83.3% of SNP1, 42.2% of SNP2 and 61.2% of SNP3. Among the hidden services that are absent in just one of the three data sets, especially notable are the 76 hidden services that reappeared in SNP3 after they disappeared during SNP2.

To provide a better picture of the complexity of Tor websites, for each and every hidden service, we proceeded as follows: *i*) we reconstructed the whole tree-structure of subdomains and pages; *ii*) we computed the total number of characters and the total number of hyperlinks (*i.e.*, number of *href*s in the HTML source). Figure 1 shows the statistical distribution of tree heights for the three snapshots and the distribution of tree height variations across different snapshots (for hidden services present in, at least, two snapshots). The trees are generally very short and do not vary remarkably over time, yet exceptions exist with variations comparable to the maximum "size" of a hidden service. The char count is generally variable, whereas services with 0 hyperlinks are predominant. A significant number of hidden services has one hyperlink every 20 to 200 chars (*i.e.*, from $\approx$ 3 words up to $\approx$ 2 sentences). In the following sections we rely on the ratio of number of hyperlinks over number of characters (*links-to-char ratio*, or LCRatio) to assess whether hidden services that are central in the Tor Web graph are indeed just link directories or not. It is worth noting that, of the 10685 hidden services reached in all three snapshots, only $\approx$ 65% had a constant tree height and only $\approx$ 43% had a constant char count across all snapshots. Automatically detecting hidden services that stay durably online but with different names (*e.g.*, to prevent being tracked down) thus requires manual work that lies beyond the scope of the present paper.

For contents analysis we rely on the DUTA dataset, the widest publicly available thematic dataset for Tor, consisting of a three-layer classification of 10250 hidden services [1, 2]. Albeit the DUTA classification does not cover our dataset entirely, the percentage of HSs of our snapshots contained in the DUTA dataset is significant: for instance, $\approx$ 49.5% of the fully persistent HSs found in all three snapshots, and $\approx$ 85% of the 200 HSs having most hyperlinks to other HSs, have a DUTA tag. In addition, the DUTA dataset has the

Table 1 Services persistence over time; in total we reached almost 30000 different hidden services

Snapshot	Number of HS	Percentage of HS	Persistence Type
SNP1 and SNP2 and SNP3	10685	36.25%	Fully persistent
SNP1 and SNP2	1612	5.47%	Partially persistent
SNP2 and SNP3	3066	10.40%	Partially persistent
SNP1 and SNP3	76	0.26%	Reappeared
SNP1 only	456	1.55%	Seen once
SNP2 only	9945	33.74%	Seen once
SNP3 only	3633	12.33%	Seen once



(a) Distribution of tree heights (b) Distribution of tree height variations

Fig. 1 Distribution of tree heights in the three snapshots (**a**) and distribution of tree height variations across different snapshots for hidden services present in at least two snapshots (**b**)

undeniable advantage of being manually tagged – by choosing it rather than carrying out a fresh new classification of our dataset, we trade coverage for accuracy.

The DUTA dataset provides a two-layer thematic classification plus a language tag for each service. The thematic classes are further categorized as "Normal", "Suspicious" or "Unknown". The "Unknown" category only includes classes that correspond to services whose nature could not be established: "Empty", "Locked" or "Down". Due to the limited information provided by these tags, we ignore all "Unknown" services in the following. For certain first layer classes (*e.g.*, "Marketplace") that can be both "Suspicious" and "Normal", the second layer is exactly used to tell apart "Legal" and "Illegal" content. We consider the second layer for this purpose only, thus obtaining the customized version of the DUTA thematic classification reported in Table 2.

3 Methods

3.1 Graph construction

From each of the three WARC [1] files obtained from the scraping procedures we extracted two graphs: a *Directed* Service Graph (DSG) and an *Undirected* Service Graph (USG). As detailed in [12], a vertex of these graphs represents the set of pages belonging to a hidden service. In the DSG a directed edge is drawn from hidden service HS1 to HS2 if any page in HS1 contains, at least, a hypertextual link to any page in HS2[2]. The directed graphs obtained from the three snapshots are denoted DSG1, DSG2, and DSG3, respectively. In the USG, instead, an undirected edge connects hidden services HS1 and HS2 if they are *mutually* connected in the corresponding DSG, that is, if there exists at least one page in HS1 linking any page in HS2 *and* at least one page in HS2 linking any page in HS1. More formally an edge $(u, v) \in E_{USG}$ iff $(u, v) \in E_{DSG}$ and $(v, u) \in E_{DSG}$, Figure 2 shows an example of construction of a DSG and a USG. When we consider just mutual connections, a vast majority of vertices remains isolated. These are ignored in the following since they

[1] Web ARChive https://www.iso.org/obp/ui/#iso:std:iso:28500:ed-2:v1:en

[2] Edges from/to the surface web have been ignored.

Table 2 The content-based classification used in this paper

Class name by type	
Normal	Suspicious
Art	Counterfeit Credit-Cards
Casino	Counterfeit Money
Cryptocurrency	Counterfeit Personal-Identification
Forum (Legal)	Cryptolocker
Hosting	Drugs
Library	Forum (Illegal)
Marketplace (Legal)	Fraud
Personal	Hacking
Politics	Human-Trafficking
Religion	Leaked-Data
Services (Legal)	Marketplace (Illegal)
Social-Network	Porno
	Services (Illegal)
	Violence

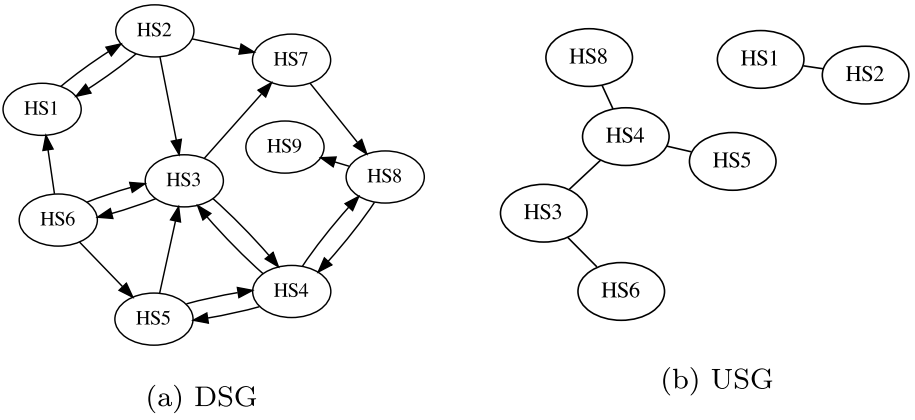


Fig. 2 A toy example showing a *Directed* Service Graph (DSG) and an *Undirected* Service Graph (USG). The USG is built from the DSG by keeping only *mutually* connected hidden services. We consider edge-induced graphs, thus isolated vertices are ignored

convey no structural information. In other words, we consider edge-induced graphs. The undirected graphs obtained from the three snapshots are denoted USG1, USG2, and USG3 respectively.

Since the snapshot graphs are inevitably conditioned by the effect of scraping a reputedly volatile network, we also consider the edge-induced intersection and union of the aforementioned graphs. Precisely, we denote DSGI the graph induced by the edge set $E_{DSGI} = E_{DSG1} \cap E_{DSG2} \cap E_{DSG3}$ and DSGU the graph induced by the edge set $E_{DSGU} = E_{DSG1} \cup E_{DSG2} \cup E_{DSG3}$. Analogously,

USGI is induced by the edge set $E_{USGI} = E_{USG1} \cap E_{USG2} \cap E_{USG3}$ and USGU is induced by the edge set $E_{USGU} = E_{USG1} \cup E_{USG2} \cup E_{USG3}$.

We do not preserve multi-edges in order to allow a direct comparison with most previous work on other web and social/complex networks. However, in both directed and undirected graphs, we store the information about the number of links that have been "flattened" onto an edge as a *weight* attribute assigned to that edge – taking the minimum available weight for edges of our intersection graph and the maximum for the union. We interpret the edge weight as a measure of connection strength that does *not* alter distances but expresses endorsement/trust and quantifies the likelihood that a *random web surfer* [33] travels on that edge.

3.2 Graph Analysis

In line with previous work on Web and social graphs [11, 19, 21, 26], we analyze the Tor Web graph through a set of structural measures and statistics, including a bow-tie decomposition of the directed graphs, global and local (*i.e.*, vertex-level) metrics, and modularity-based clustering. The main graph-based notions and definitions are reported in the following, while graph-related symbols used throughout the paper are reported in Table 3.

Bow-Tie decomposition In a directed graph, two vertices u and v are strongly connected if there exists a path from u to v and a path from v to u. Strong connectedness defines equivalence classes called strongly connected components. A common way to characterize a directed graph consists in partitioning its vertices based on whether and how they are connected to the largest strongly connected component of the graph. This "bow-tie" decomposition [26] consists of six mutually disjoint classes, defined as follows: (i) a vertex v is in LSCC if v belongs to the largest strongly connected component; (ii) v is in IN if v is not in LSCC and there is a path from v to LSCC; (iii) v is in OUT if v is not in LSCC and there is a path from LSCC to v; (iv) v is in TUBES if v is not in any of the previous sets and there is a path from IN to v and a path from v to OUT; (v) v is in TENDRILS if v is not in any of the previous sets and there is either a path from IN to v or a path from v to OUT, but not both; otherwise, (vi) v is in DISCONNECTED.

Global metrics To characterize our ten graphs we resort to well-known metrics, summarized in Table 4. Most of these metrics have a straightforward definition. Let us just mention that: in directed graphs, following Newman's original definition [30], the assortativity ρ measures the correlation between a node's out-degree and the adjacent nodes' respective in-degree; in undirected graphs, ρ measures the correlation between a node's degree and the degree of its adjacent nodes; the global efficiency E_{glo} is the average of inverse path lengths; in directed graphs, the transitivity T measures how often vertices

Table 3 Basic graph notations and definitions used throughout the 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: $M = \|E\|$
σ_{vu}	Number of shortest paths from v to u
$\sigma_{vu}(t)$	Number of shortest paths from v to u including t
$\text{dist}(v, u)$	Shortest path length from v to u

Table 4 Global metrics notations and definitions

Symbol	Definition
Global metrics valid for both DSG and USG	
$\langle \deg \rangle$	Average (in-/out-) degree
ρ	Assortativity: see (26) in [30]
d	Diameter: $\max_{v \in V} \max_{u \in V} \text{dist}(v, u)$
$\langle \text{dist} \rangle$	Average shortest path length
E_{glo}	Global efficiency: $\frac{1}{N(N-1)} \sum_{u \neq v \in V} \frac{1}{d(u,v)}$
Global metrics valid for DSG only	
Δ_{in}/N	Normalized maximum in-degree
Δ_{out}/N	Normalized maximum out-degree
Cen_{out}	Out-degree centralization: $\frac{N * \Delta_{out} - \sum_{v \in V} \deg_{out}(v)}{(N-1)^2}$
T	Global transitivity: $\frac{\#(u,v,w):\ u \to v \wedge u \to w \wedge (v \to w \vee w \to v)}{\#(u,v,w):\ u \to v \wedge u \to w}$
Global metrics valid for USG only	
Δ/N	Normalized maximum degree
Cen	Degree centralization: $\frac{N * \Delta - \sum_{v \in V} \deg(v)}{(N-1)(N-2)}$
C	Global clustering coefficient: $\frac{\#\text{ closed triplets}}{\#\text{ all triplets}}$

that are adjacent to the same vertex are connected in, at least, one direction; the clustering coefficient C is the transitivity in undirected graph, defined as the ratio of closed triplets over total number of triplets. Many of the metrics from Table 4 are undefined for disconnected graphs, or may provide misleading results when evaluated over multiple isolated components. To make up for it and allow for a fair comparison, we only consider the giant (weakly) connected component of all disconnected graphs. It is worth mentioning that the three Directed Service Graphs (DSGs), and therefore their union DSGU, consist of a single weakly connected component. On the contrary, all Undirected Service Graphs (USGs) are weakly disconnected graphs. DSGI is also disconnected, albeit only two hidden services – violet77pvqdmsiy.onion and typefacew3ijwkgg.onion – are isolated from the rest and connected by an edge. We instead consider the graphs in their entirety for other types of analysis.

Correlation analysis of centrality metrics We perform a correlation analysis of several local structural properties to the purpose of sorting out the possible roles of a service in the network. We rely on Spearman's rank correlation coefficient – rather than the widely used Pearson's – for a number of reasons: (i) we are neither especially interested in verifying linear dependence, nor we do expect to find it; (ii) we argue that not all the considered metrics yield a clearly defined interval scale – while they apparently provide a ordinal scale; (iii) when either of the two distributions of interest has a long tail, Spearman's is usually preferable because the rank transformation compensates for asymmetries in the data; and (iv) recent work [27] showed that Pearson's may have pathological behaviors in large scale-free networks. The considered metrics[3] are shown in Table 5. In words: the betweenness of v measures the ratio of shortest paths that pass through v; the closeness of v is the inverse of the average distance of v from all other vertices; the pagerank of v measures the likelihood that a random web surfer ultimately lands on v; the authscore and hubscore of v, jointly computed by the HITS algorithm [25], respectively measure how easy it is to reach v from

[3] Beware that some of these metrics are only defined for directed graphs.

Table 5 Local metrics notations and definitions

Short name	Full name and definition
betweenness	Betweenness centrality: $BC(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$
closeness	Closeness centrality: $CC(v) = \frac{N-1}{\sum_{u \in V} d(u,v)}$
pagerank	PageRank: see [17]
authscore	Authority score: see [24]
hubscore	Hub score: see [24]
efficiency	Efficiency: $E(v) = \frac{1}{deg(v)(deg(v)-1)} \sum_{u \neq w:\ v \to u \wedge v \to w} \frac{1}{d(u,w)}$
transitivity	Transitivity: $T(v) = \frac{\#(u,w):\ v \to u \wedge v \to w \wedge (u \to w \vee w \to u)}{\#(u,w):\ v \to u \wedge v \to w}$
eccentricity	Eccentricity: $\epsilon(v) = \max_{u \in V} d(v,u)$
LCRatio	Links-to-chars ratio, see Section 2

a central vertex or to reach a central vertex from v; the efficiency of v is the average inverse distance of v from all other vertices; the transitivity of v is the ratio of pairs of neighbors of v which are themselves adjacent; the eccentricity of v is the maximum distance of any other vertex from v; the LCRatio of v is not a graph-based metrics, but we defined it as the ratio of number of hyperlinks over number of characters in the text extracted from the HS associated to v.

Degree distribution We perform a log-normal and a power-law fit of the degree distribution of all graphs using the statistical methods developed in [14], relying on the implementation provided by the POWERLAW python package [3]. A log-normal distribution may be a better fit of degree distributions in many complex networks [28], and a recent work suggests that a log-normal distribution may emerge from the combination of preferential attachment and growth [35]. Nevertheless, using a power-law fit is standard practice in the study of long-tailed distributions and allows direct comparison with previous works. It is worth specifying that POWERLAW autonomously finds a lower-bound $k_{\min}$ for degrees to be fitted. In our case, even if $k_{\min}$ is much less than the maximum degree, all values greater than $k_{\min}$ account for just a small percentage of the whole graph. However, we believe this should not prevent from taking these fits seriously into consideration: the tail of the distribution *de facto* describes the central part of the graph that actually has a meaningful structure – as opposed to the bulk of the distribution mostly depicting vertices with out-degree 0 (83% to 95% of the graph according to the specific DSG considered) and/or in-degree 1 (17% to 43%). The procedure by which we calculate the reach of the most important hubs of each network is the following: taking into account just the giant component, we *i*) sort the hidden services by degree (out-degree in the DSGs); *ii*) compute the cumulative percentage of the giant component that is at distance one from one of the first i hubs, for $i \in \{1, \ldots, 25\}$.

Community structure To extract a community structure for our graphs we rely on the well-known Louvain algorithm [9], based on modularity maximization. As often done in the literature [19], we consider edge weights to make it harder to break an edge corresponding to a hyperlink that appears several times in the dataset. To compare the clusters emerged across different graphs, we consider how common vertices are grouped in each graph using the well-known Adjusted Mutual Information (AMI) to measure the similarity of two partitions. The AMI of two partitions is 1 if the two partitions are identical, it is 0 if the mutual information of the two partitions is the expected mutual information of two

random partitions, and it is negative if the mutual information of the two partitions is worse than the expected one. Since a single label from Table 2 is assigned to each service, the DUTA classification naturally induces three hard partitions, denoted "duta" (the individual classes), "duta type" (the macro categories "Normal" and "Suspicious") and "lang" (the language) in the following. For the set of hidden services that our graphs share with the DUTA dataset, we can assess the coherence of topic-based and modularity-based clustering by computing the AMI of "duta", "duta type" and "lang" with respect to the Louvain's clusters.

3.3 Topological features for content-based classification

To measure the information gain provided by topological vertex properties with respect to content-based classification, we proceed as follows:

- For each DUTA category C, we consider the dummy variable X_C that indicates whether a randomly picked service belongs to the considered category.
- We let each metrics m induce a probability distribution P_m over the set of all services, in such a way that the probability of selecting a HS is proportional to the value of that metrics for that service.
- To measure the importance of knowing a metrics m with respect to a specific category C, we compare the distribution of X_C under two different assumptions: that the HSs are drawn based on P_C and that they are drawn uniformly at random – the latter meaning that $\Pr[X_C = 1]$ is the overall prevalence of C in the graph.
- As a measure of information gain, we use the Kullback-Leibler divergence. The KL divergence lies in $[0, +\infty]$, and it is 0 if the two distributions coincide.

4 Results and discussion

Hereafter, we summarize and discuss our main findings; additional explanations, statistics and figures are available in the Appendices. Since we monitored Tor over a sufficient time span, our analysis is robust under fluctuations of the results obtained for different snapshots. The union and intersection graphs, in particular, capture most of the features of the snapshots, reflecting in different ways some of their specific characteristics. We will therefore often focus on such graphs to provide a clear and synthetic overview of the results.

The bow-tie decomposition of the DSGs is reported and compared with previous work in Table 6. In general agreement with [20], we found that the Tor Web has a radically different structure with respect to the WWW, except, in part, for the DSGI graph where all components are non-empty. The Tor Web consists of just a very small LSCC and a much larger OUT component, albeit the share of the LSCC in the total size of the graph may be heavily influenced by the volatility of the network.

At a first sight, the Tor Web, seen as a directed graph, seems to show the key features of a *small world* network: the transitivity T is one order of magnitude greater than $\langle \deg \rangle / N$, which is the expected transitivity in a comparable random graph; the distance between any two connected nodes is approximately logarithmic in N, as in most social and web graphs

Table 6 Bow-Tie structure

Graph	Component					
	LSCC	IN	OUT	TUBES	TENDRILS	DISCONNECTED
	22.3M	3.3M	13.3M	17K	514K	3.5M
WWW from [26]	51.94%	7.65%	30.98%	0.04%	1.2%	8.2%
Tor from [20]	297	0	6881	0	0	0
	4.14%	0.0%	95.86%	0.0%	0.0%	0.0%
	466	0	12363	0	0	0
DSG1	3.63%	0.0%	96.37%	0.0%	0.0%	0.0%
DSG2	820	0	24488	0	0	0
	3.24%	0.0%	96.76%	0.0%	0.0%	0.0%
	2371	0	15089	0	0	0
DSG3	13.58%	0.0%	86.42%	0.0%	0.0%	0.0%
DSGI	169	9	7415	1	74	1
	2.2%	0.12%	96.69%	0.01%	0.97%	0.01%
	3062	0	26411	0	0	0
DSGU	10.39%	0.0%	89.61%	0.0%	0.0%	0.0%

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)

(see Appendix 1). A typical small world network, however, should be efficient, while the Tor Web has a very low global efficiency (E_{glo}), which is computed assigning infinite distance to non-connected vertex pairs.

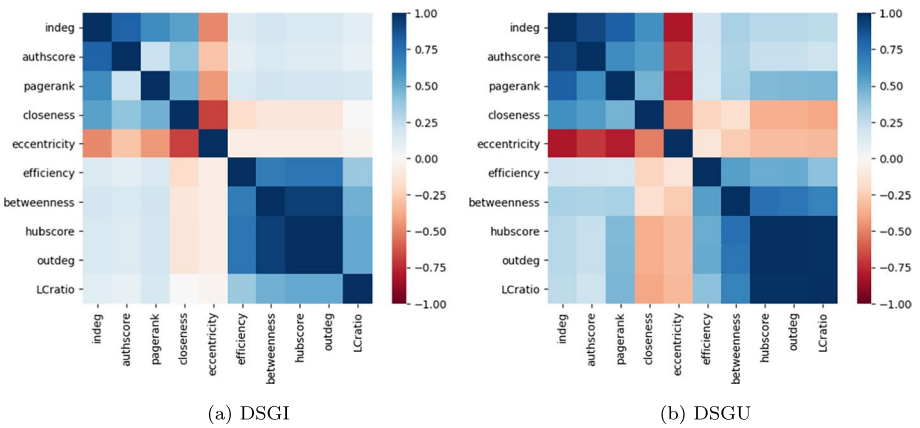
The emerging structure entails that most resources of potential interest for Tor users are not easy to reach. The only way to effectively browse this network is to find a HS that belongs to the LSCC (possibly, through a link from the surface Web) and, from there, look for a path to the resource of interest. Navigation in the network is thus mostly tied to paths that connect the tiny LSCC to the many peripheral nodes, whereas the inverse route is precluded. This shows that the user experience in Tor is quite different from that of the WWW, and supports the general perception that most Tor users do not actually browse the Tor Web, but already know the onion url they want to visit.

The small-world effect that we observe in the DSGs is not visible in the undirected version of the Tor Web graphs, which have far fewer vertices but a comparable, or even greater, average distance and diameter than their directed counterparts. Most of the paths in the USGs, however, have length close to $\langle \text{dist} \rangle$, so that $E_{glo} \approx 1/\langle \text{dist} \rangle$ and the graphs are thus much more efficient than the DSGs. The clustering coefficient C is large in the USGI, but drops to $\approx \langle \text{deg} \rangle / N$ in the USGU due to the presence of a huge hub in USG3.

To gain insights into the local properties of the network, we performed a correlation analysis of several centrality metrics (see Appendix 2 for additional details). In Figure 3 we see that in the DSGs the central vertices can be broadly categorized in two groups. On the one hand, HSs having large in-degree, authscore, pagerank and closeness are those that provide the most valuable content: they attract connections from all other HSs, including "important" ones and are, on average, easier to reach. On the other hand, HSs having large out-degree, hubscore, betweenness and efficiency provide significant contribution to information flows and are at the center of highly clustered regions. The fact that the LCRatio correlates with the latter set of metrics suggests that these hubs are mostly link directories or similar Web services.

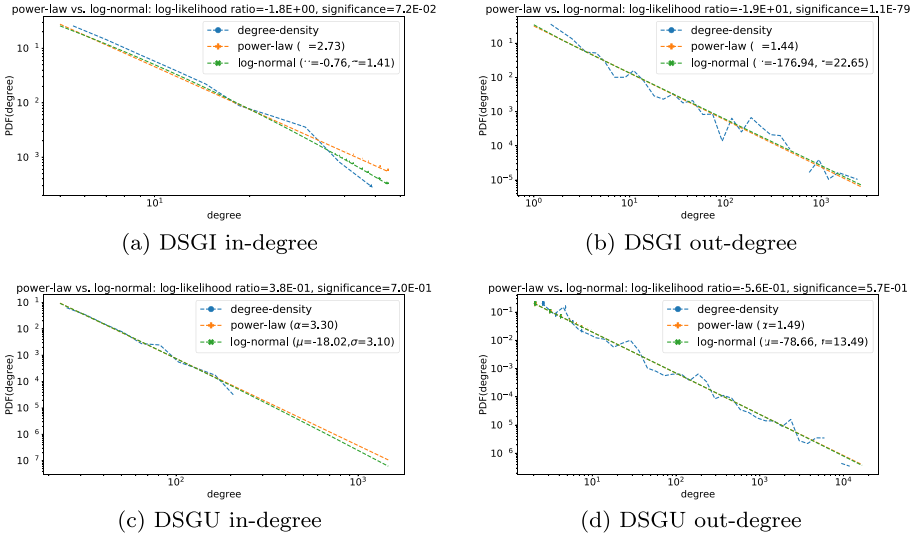
Supported by the correlation analysis, we then focused on the degree sequences (in- and out-degree for the DSGs) to gain information on the hierarchical organization of the network. The tails of the distributions, in particular, describe the central part of the graph having a meaningful structure, whereas the bulk of the distribution mostly depicts peripheral HSs with very low degree. We performed both a log-normal and a power-law fit of the degree distributions (see Appendix 3 for details). While the former is slightly more accurate, power-law fits are widely used in the literature and looking at the α exponent of the power-law is a straightforward way to classify the Tor Web graph with respect to the vast body of work on complex networks.

The in- and out-degree distribution of the DSGI and DSGU are shown in Figure 4. The value of α obtained for the out-degree distribution lies consistently around 1.5 for all directed graphs. This may be interpreted as the emergence of some level of self-organization: the choice of how many links to include in its web pages, arguably taken in full autonomy by each HS, makes the network resilient and facilitates its navigability. While the low value of α obtained for the out-degree distribution shows that hubs are quite common in Tor, the strong out-degree centralization Δ_{out} signifies that some of these hubs are especially large if compared with the others. In the DSGU, more than 90% of the graph is in fact at distance 1 from (at least) one of the top 6 hubs; in the DSGI, which is much less centralized, more than 90% of the graph is still at distance 1 from (at least) one of the top 23 hubs (see Appendix 3).



(a) DSGI (b) DSGU

Fig. 3 Spearman's rank correlation coefficient between the considered local metrics for the DSGI and DSGU



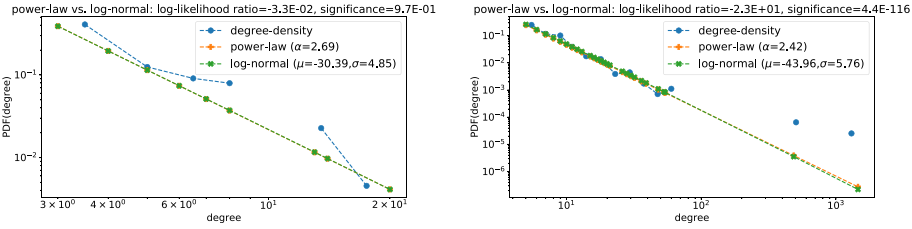
(a) DSGI in-degree (b) DSGI out-degree

(c) DSGU in-degree (d) DSGU out-degree

Fig. 4 The in- and out-degree distribution for the DSGI and DSGU

All DSGs are *disassortative*, meaning that most of the HSs linked by a hub have a low in-degree. This means that the neighborhoods of different hubs are, at least, partially disjoint. As we have seen, the in-degree correlates with a set of metrics that express the authority and measure the ease of reach of a service. For the in-degree, α lies around the threshold 3 that is known to control the variance of the distribution, with $\alpha \approx 2.7$ in the DSGI and $\alpha \approx 3.3$ in the DSGU. This indicates that even authorities have a moderate in-degree and that, to match up with the out-degree, there are many HSs with a very low in-degree that may become almost impossible to reach due to minimal changes in the Tor link connectivity. Combined with the disassortativity, and contrary to what the out-degree distribution may suggest, this means that access to valuable information is barely granted in Tor.

The degree distribution of the USGs mostly follows a power-law with α exponent ≈ 2.5, closer to the value typically found in social networks, as visible in Figure 5. Mutual connections seem to represent the backbone of the social structure of the Tor Web graph, as also confirmed by a comparison of the distribution of DUTA topics in the DSGs and USGs (see Figure 6). While the DSGI and the DSGU follow the original DUTA distribution quite



(a) USGI (b) USGU

Fig. 5 The degree distribution for the USGI and USGU

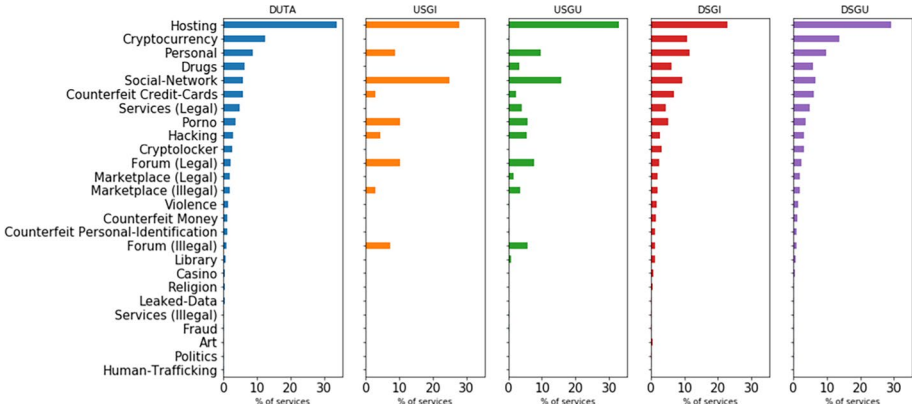


Fig. 6 Distribution of tags from Table 2 in the DUTA dataset and in the four considered Tor Web graphs

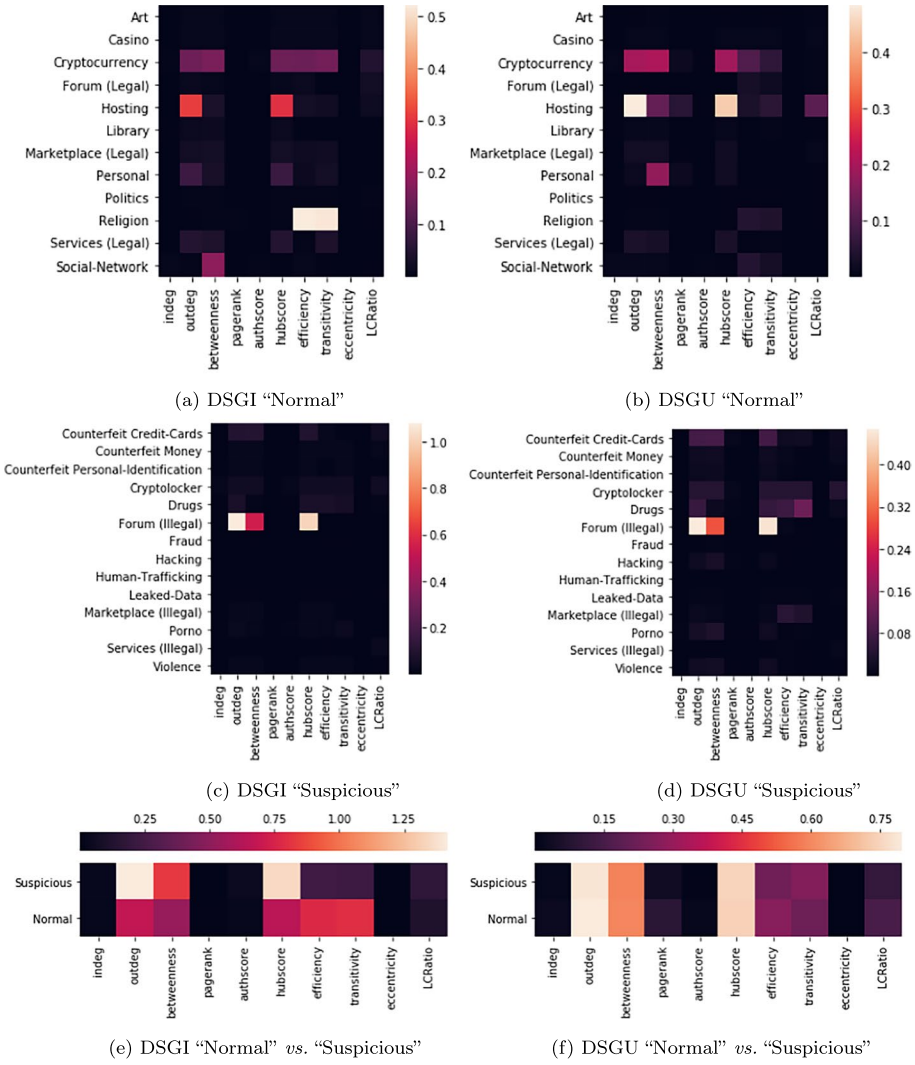
closely, the thematic tag distributions in the USGs show an increased prevalence of classes of HSs related to sociality in a broad sense, such as "Social Network" or "Forum". This type of HSs are keener to link to other similar HSs, thus favoring the emergence of mutual links. More generally, depending on the tag, HSs could instead compete and, hence, not connect with one another. This could explain why common classes such as "Cryptocurrency" or "Drugs" are entirely missing or barely present in the USGI and in the USGU.

A few outliers in the degree sequence of the USGU show that the inferred scale-free distribution cannot fully explain the organization of mutual links in Tor. The combined neighborhoods of the two most central HSs, in particular, cover almost 90% of the USGU. The USGI, instead, is much less centralized. As all networks are again *disassortative*, we know that also in USGs hubs are more likely connected with peripheral nodes.

We inferred the community structure of our graphs through modularity-based clustering and measured the similarity of the clusters obtained for different graphs computing the Adjusted Mutual Information (AMI) on the clusters projected on the set of common vertices. The community structure of the DSGs is very similar, in terms of number and size of the clusters, and reasonably consistent, taking into consideration the volatility of the network, with AMI ≈ 0.5 for all combinations (see Appendix 4). In any case, the apparent significance of the obtained clusters does not respond to a thematic homogeneity: for the set of HSs that our graphs share with the DUTA dataset, the coherence of topic-based and modularity-based clustering resulted in an AMI score ≈ 0.

While USGs have a more heterogeneous community structure, common vertices are clustered in an extremely stable way in the USGs, meaning that the existence of a mutual link is – as expected – a stronger indicator of the similarity between two services. We also see that the union graphs DSGU and USGU, *i.e.*, the graphs based on all collected data, are those whose community structure is less influenced by switching from the directed to the undirected graph. In some sense, this means that the clustering obtained for DSGU can be reasonably considered as an extension of the very meaningful partition obtained for USGU.

To assess whether computing graph-based centrality metrics provides any advantage to the purpose of inferring the thematic tag of a HS, we proceeded as described in Section 3.3. In Figure 7 we show the measured information gain for the DSGI and DSGU, separately considering "Normal" classes, "Suspicious" classes and their aggregate. The scenario for the other DSGs is almost identical and thus omitted, whereas the USGs were



(a) DSGI "Normal"

(b) DSGU "Normal"

(c) DSGI "Suspicious"

(d) DSGU "Suspicious"

(e) DSGI "Normal" *vs.* "Suspicious"

(f) DSGU "Normal" *vs.* "Suspicious"

Fig. 7 The information gain provided by different metrics with respect to DUTA classes and macro categories

not considered because their limited size affects the statistical relevance of this method. Generally speaking, most of the metrics appear to be uninformative with respect to content-based categories, *i.e.*, the probability of finding a service of a specific class does not increase or decrease significantly when we select the service with probability proportional to most of its topological properties. However, there are a few remarkable exceptions: (i) the out-degree and the hubscore are especially informative about hosting services and illegal forums; (ii) services discussing religion topics are highlighted by their efficiency and transitivity, arguably because they tend to strongly cluster together; (iii) in the DSGU, the transitivity is also somewhat informative of services that focus on drugs, whereas the LCRatio is associated with hosting services, even though not as much as one could expect.

These class-level information gains are only partially able to explain the notable improvement that many metrics instead seem to provide to the goal of telling apart, more in general, "Suspicious" and "Normal" services. This opens new perspectives towards the design of classifiers that make use of topological features instead of text analysis.

5 Conclusion

In this paper, we presented an in depth investigation of the key features of the Tor Web graph, providing a clear view on its topology and on how the topology is affected by the volatility of the network, inferring on the latent patterns of interactions among Tor users, and assessing whether graph metrics can be used to expose the thematic organization of the network. The Tor Web is composed of a large percentage of volatile hidden services and of mostly persistent hubs that are critical for the graph connectivity. The volatility of peripheral nodes does not heavily influence the global structure of the Tor Web graph, which consists of a small strongly connected component from which the remainder of the network can be reached in just a few steps. Albeit a small world effect can be observed, the Tor Web has a very low global efficiency and most resources of potential interest for Tor users are not easy to reach. The graph seems to possess a meaningful and stable community structure, not related to the thematic organization of the network, which is especially visible when only mutual connections are considered. The subgraph induced by mutual connections comprises just a tiny fraction of the nodes and includes a major presence of topics related to sociality in a broad sense. Considering a class-level categorization, most of the applied topological metrics appear to be scarcely informative with respect to the hidden services' content. Nevertheless, some metrics seem to provide a notable improvement in the goal of telling apart "Suspicious" from "Normal" services.

We are used to consider the Web and online social networks as systems in which we can find or disseminate information. The Tor Web does not seem to be based on these two cornerstones: it is inefficient in spreading information and difficult to navigate. If compared with most real world complex networks, it has a fairly simple and asymmetric structure that is reflected in its navigation being facilitated only in one direction: users select a starting out-hub and then they move looking for the website of interest. Peripheral nodes, once reached, usually do not provide any possibility to go back and navigate in other directions. The number of hops required to reach a node, when possible, remains limited, but the overall structure is quite different compared to a typical *small world* network. As a consequence of these topological features, Tor provides a very different user experience from that of the WWW and online social media.

Future efforts will be devoted to widen the scope and the depth of the analysis. Any study of the dynamics on and of Tor would benefit from monitoring the Tor Web consistently over a long time range and possibly measuring the influence of exogenous factors (*e.g.* changes in the legislation or breaking news from the real-world) on the Tor Web organization. More generally, crawling specific areas of the surface Web (*e.g.*, forums on Reddit or public groups on Whatsapp or Telegram) may lead to onion urls that could not be found scraping Tor itself. This could either confirm that the majority of Tor's HSs are isolated from the subset of the Tor Web having a network structure, or, to the contrary, reveal a more complex system composed of multiple portions of the Tor Web connected through a layer of surface websites.

Table 7 Outcomes of the three crawling processes

Crawl	End Date	# records per response type				
		2xx	3xx[a]	4xx[b]	5xx[b]	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] A status code 3xx is related to Web redirection (https://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html).

[b] Status codes 4xx and 5xx are error codes.

Appendix A: Data collection

To collect data from the Tor Web we used a customized crawler fed with a list of seeds. 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"[4]), 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. As result, we obtained three different "snapshots" of the Tor Web, denoted SNP1, SNP2, and SNP3, respectively. Table 7 describes our datasets the composition of which is comparable to similar studies in the Literature [36]. Yet, if we refer to the statistics provided by the Tor Project for the corresponding time window[5], 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 [20] and by hidden services that do not host websites.

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[6] [23], Heritrix[7] [29] and BUbiNG [10]. 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 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. 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

[4] Here, the meaning of "random" depends on the choice of a distribution over the set of all possible partitions [38]

[5] wikitjerrta4qgz4.onion

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

[7] http://nutch.apache.org

Table 8 Global metrics for the directed service graphs

	DSG1	DSG2	DSG3	DSGI	DSGU
N	12829	25308	17460	7669	29473
M	72556	113014	103402	28913	187415
$\langle \text{dist} \rangle$	3.793	4.96	3.665	3.983	3.821
d	10	12	10	10	9
E_{glo}	0.011	0.008	0.041	0.007	0.029
$\langle \text{deg} \rangle$	5.656	4.466	5.922	3.77	6.359
$\frac{\Delta_{\text{in}}}{N}$	0.016	0.01	0.084	0.007	0.05
$\frac{\Delta_{\text{out}}}{N}$	0.437	0.508	0.611	0.348	0.566
ρ	−0.319	−0.327	−0.162	−0.374	−0.168
Cen_{out}	0.436	0.508	0.61	0.348	0.566
T	0.004	0.002	0.002	0.004	0.002

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.

Appendix B: Extended results

B. 1 Global metrics

The global metrics for our DSGs are reported in Table 8. The variance in the sizes N and M of the three snapshots is consistent with publicly available aggregated statistics[8], as already discussed in [6]. The values of Δ_{out}/N and Δ_{in}/N say that the main out-hubs reach 35% to 61% of the network, whereas no equivalently prominent in-hubs exist. The values of Cen_{out} show that these main out-hubs have a prominent role in the graphs' connectivity. However, the greatest of such hubs emerges in the largest graphs, and Δ_{in}/N and Δ_{out}/N are comparably smaller in DSGI with respect to the snapshots, suggesting that the degree of such stable hubs is heavily influenced by the non-persistent nodes. All networks are *disassortative*, meaning that links are more likely to connect high-out-degree nodes to low-in-degree nodes, or low-out-degree nodes to high-in-degree nodes. By comparing the transitivity T with $\langle \text{deg} \rangle / N$, we see that in our graphs two vertices that are adjacent to the same vertex are connected (in, at least, one direction) significantly more often than in a random graph. The diameter d and the average path length $\langle \text{dist} \rangle$ are approximately logarithmic in N. The emergence of these two properties is usually denoted *small world* effect. Another quantity often used to quantify small world behavior in networks is the global efficiency E_{glo}, defined as the average of inverse pairwise distances, where disconnected pairs of vertices have infinite distance. Since $E_{glo} \ll 1/\langle \text{dist} \rangle$, we realize that our Tor graphs are inefficient because many pairs of nodes are disconnected and thus cannot be really considered small world networks.

Table 9 reports analogous metrics for the giant connected components of the USGs. The sizes N and M of the three snapshots are again variable, but the USGs are generally

[8] https://webarchive.jira.com/wiki/display/Heritrix

Table 9 Global metrics for the undirected service graphs

	USG1	USG2	USG3	USGI	USGU
N	208	225	2084	87	2244
M	398	467	2289	143	2685
$\langle \text{dist} \rangle$	4.301	3.941	2.707	3.939	2.881
d	15	9	10	9	11
E_{glo}	0.285	0.295	0.408	0.308	0.389
$\langle \text{deg} \rangle$	3.827	4.151	2.197	3.287	2.393
$\frac{\Delta}{N}$	0.188	0.213	0.7	0.23	0.65
ρ	−0.077	−0.121	−0.602	−0.023	−0.489
Cen	0.171	0.197	0.699	0.197	0.649
C	0.203	0.208	0.001	0.259	0.002

much smaller than the DSGs. USG3 is now the biggest, probably thanks to the existence of a huge hub that is absent in the other two snapshots, as visible from the values of Δ/N and Cen. The presence of this hub also explains why the clustering coefficient C is significantly greater than $\langle \text{deg} \rangle / N$ in USG1, USG2 and USGI, but not in USG3 and USGU. d and $\langle \text{dist} \rangle$ are comparable to or even greater than in the DSGs and $E_{glo} \approx 1/\langle \text{dist} \rangle$, meaning that most pairs are indeed at distance $\approx \langle \text{dist} \rangle$. Mutual connections thus induce a subgraph that is quite efficient but not really a *small world*. Here the assortativity ρ measures the tendency of a node to connect with others having similar degree. All networks are again *disassortative*.

B. 2 Correlation analysis of centrality metrics

Figure 8 visually shows, for the USGI and USGU, the pairwise correlation of the metrics defined in Table 5. We consider only the intersection and union graphs for the sake of clarity. In the DSGs (Figures 8a and b) the in-degree, authscore, closeness and pagerank correlate with each other, and the same happens for the out-degree, hubscore, betweenness, efficiency, transitivity and LCRatio. In other words, vertices that are authoritative are, on average, easier to reach and may not be hubs. Hubs, on the other hand, are not necessarily authoritative, they facilitate information flows and are at the center of highly clustered regions. Having a high LCRatio, the hubs are much likely link directories or similar Web services. The eccentricity is instead uncorrelated or negatively correlated with all other metrics. In the USGs almost all metrics are in general agreement, meaning that only considering mutual connections leads to a network with a well defined vertex hierarchy. There are, however, a few exceptions: in the USGI there is a lack of correlation between closeness and pagerank; in the USGU the closeness and the eccentricity "agree" with each other while they negatively correlate with all other measures; in both cases, the LCRatio is uncorrelated with all other metrics.

B. 3 Degree distribution

Figure 9 shows the distributions of the in- and out-degree for all five DSGs on a log-log scale. For DSG2, DSG3 and DSGU the fitted power-law distribution has finite variance ($\alpha > 3$), contrary to DSG1 ($\alpha \approx 2.9$) and DSGI ($\alpha \approx 2.7$). This divergent behavior

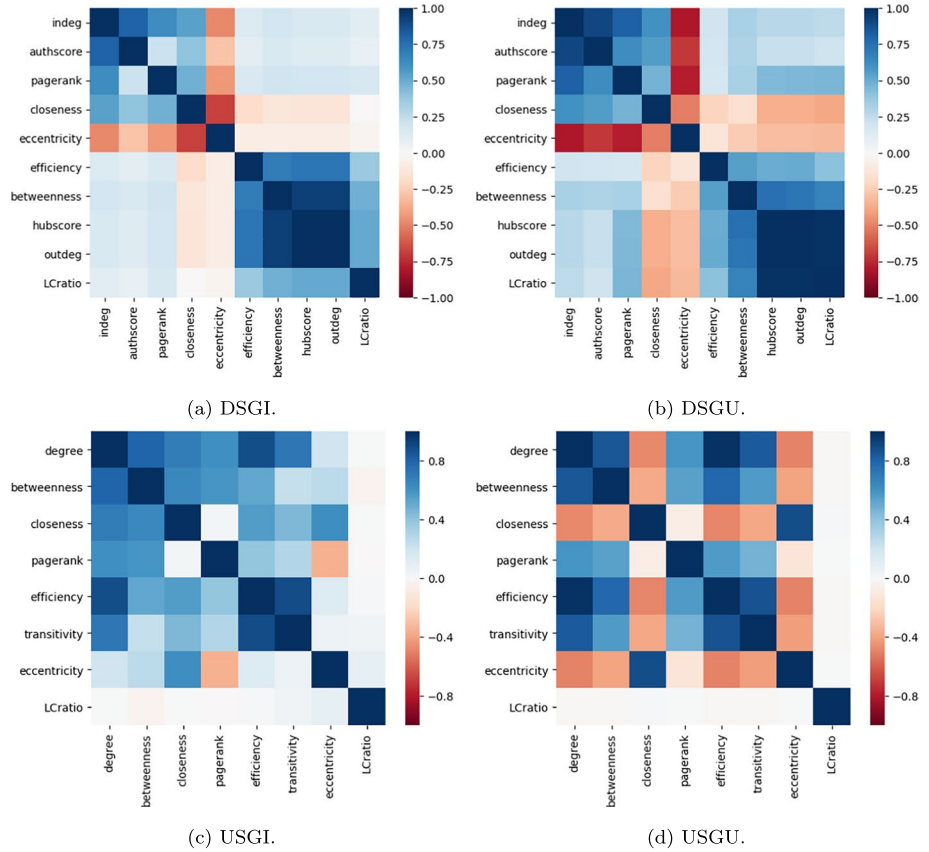
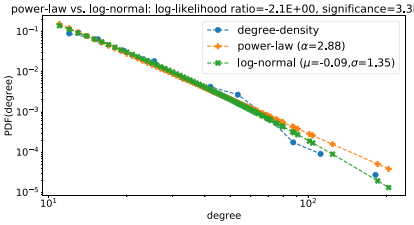


Fig. 8 Spearman's rank correlation coefficient between the considered local metrics

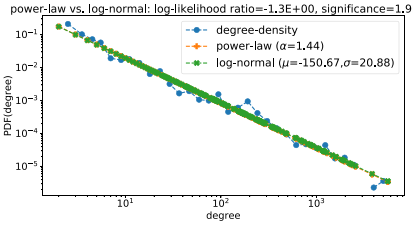
advises caution in the interpretation of this fit and suggests to mostly focus on the differences between the DSGI and the DSGU. All out-degree distributions have instead $\alpha \approx 1.5$, a very low value that reflects the existence of many large out-hubs – *i.e.*, link directories or similar web services.

In Figure 10, we report the fitted degree distribution for the USGs. We focus in particular on the USGU, which stores all available information about mutual connections between Tor's hidden services. Most of the degree sequence follows a power-law with α exponent ≈ 2.42, lower than all DSGs and typical of social networks, but huge hubs are significantly more likely to exist than in a scale-free network with such α. The plot broadly confirms the insights provided by the DSGs and shows that mutual connections are indeed the backbone of the social structure of the Tor Web graph.

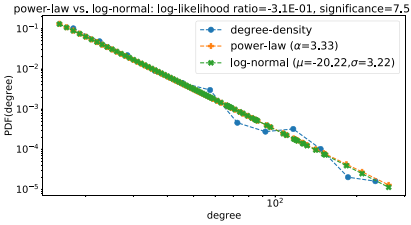
Motivated by the long tail of the degree distributions and with the purpose of gaining a better understanding of how the whole graph can be explored from just a few starting points, in Figure 11 we show how many hidden services can be reached in just one step from the top hubs. The top-6 out-degree services reach out to almost 70% of the nodes in DSGI, 80% in DSG1, and 90% in DSG2, DSG3 and DSGU, and, in all cases, the



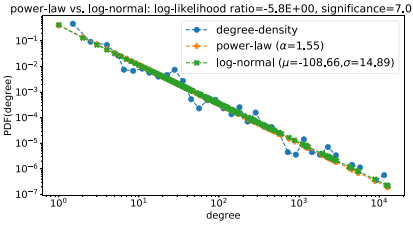
(a) DSG1 in-degree.



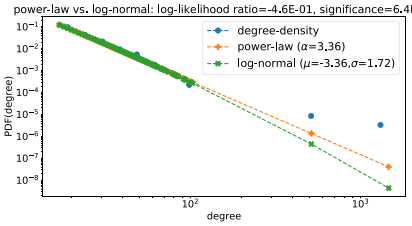
(b) DSG1 out-degree.



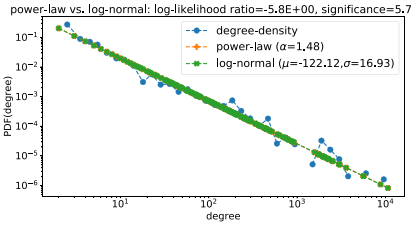
(c) DSG2 in-degree.



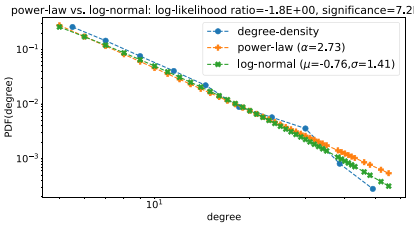
(d) DSG2 out-degree.



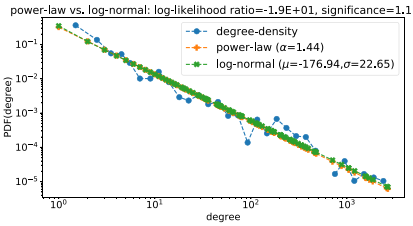
(e) DSG3 in-degree.



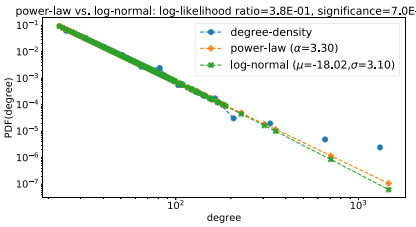
(f) DSG3 out-degree.



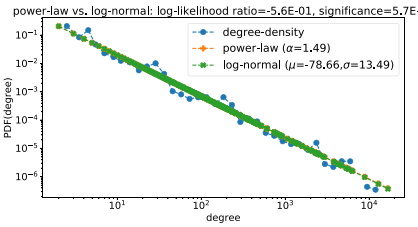
(g) DSGI in-degree.



(h) DSGI out-degree.

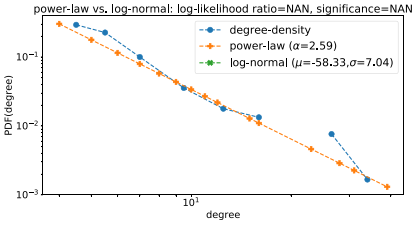


(i) DSGU in-degree.

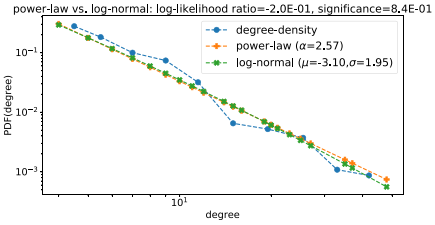


(j) DSGU out-degree.

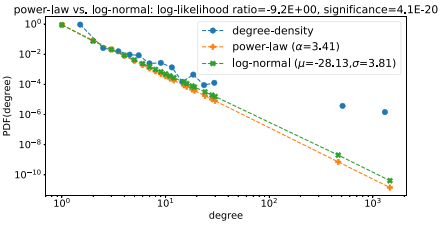
Fig. 9 The degree distribution for the DSGs



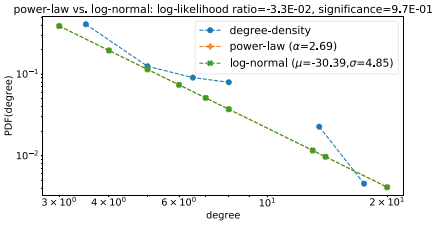
(a) USG1.



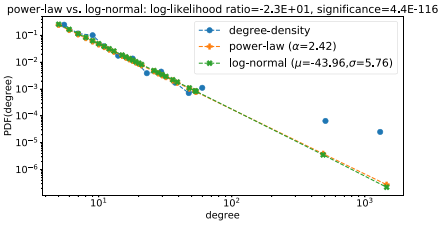
(b) USG2.



(c) USG3.

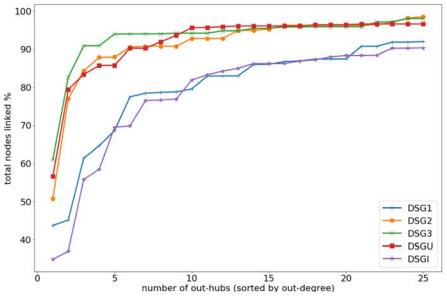


(d) USGI.

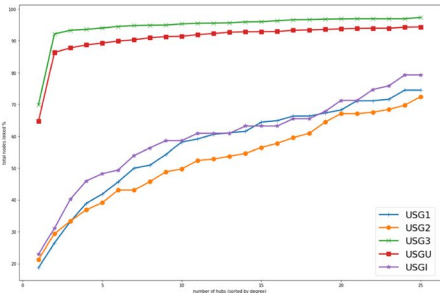


(e) USGU.

Fig. 10 The degree distribution for the USGs



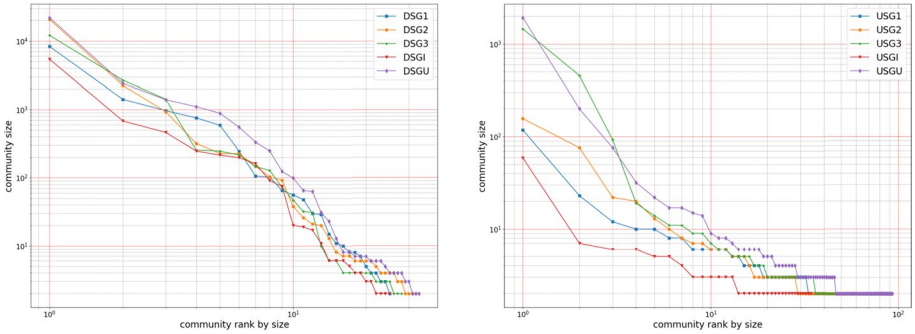
(a) DSGs.



(b) USGs.

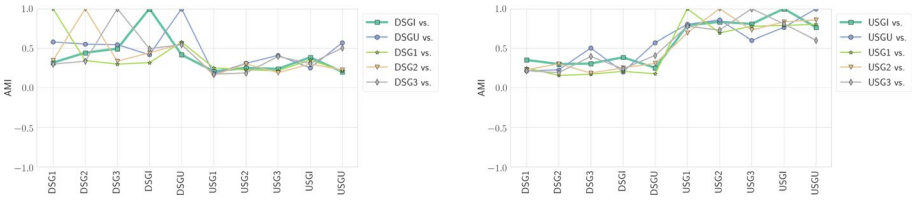
Fig. 11 Cumulative percentage of the graph linked by the top hubs

percentage quickly gets over 90% or even 95%. Among the USGs, the USG1, USG2 and USGI are much less centralized. In the USG3 – and, hence, in the USGU – the top hub alone is at distance one from more than 65% of the graph, and with just two hubs we get to



(a) DSGs. (b) USGs.

Fig. 12 The community size distribution for our Tor Web graphs



(a) DSGs *vs.* all graphs. (b) USGs *vs.* all graphs.

Fig. 13 The comparison of the partitions obtained for our Tor Web graphs

more than 85%. Again, the DSGI and DSGU are representative of two opposite behavior that may emerge from scraping the same network.

B. 4 Community structure

Figure 12 shows the distribution of cluster sizes for the DSGs (a) and the USGs (b). In Figure 13 we use the well-known Adjusted Mutual Information (AMI) to compare the clusters emerged across different graphs based on how common vertices are grouped in each graph. We recall that the AMI of two partitions is 1 if the two partitions are identical, it is 0 if the mutual information of the two partitions is the expected mutual information of two random partitions[9], and it is negative if the mutual information of the two partitions is worse than the expected one. All DSGs have a very similar structure, in terms of number and size of the clusters, and the pairwise AMI of the obtained clusters lies around 0.5. While USGs have a more heterogeneous structure, their communities are more similar, in line with the intuition that the existence of a mutual link is a stronger indicator of the similarity between two services. The only case in which a directed graph and the corresponding undirected graph have a AMI> 0.5 are the union graphs DSGU and USGU, *i.e.*, the graphs based on all collected data.

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

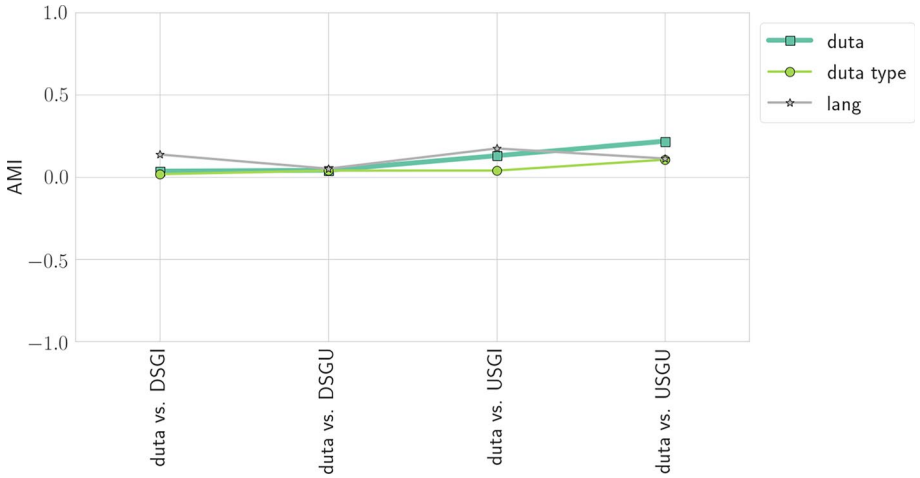


Fig. 14 The comparison of the topic-based partition induced by the DUTA dataset and the modularity-based partitions obtained through Louvain's algorithm on our graphs

To assess the coherence of topic-based and modularity-based clustering, we focused on the set of hidden services that our graphs share with the DUTA dataset and we measured the AMI of the partitions induced by the "duta", "duta type" and "lang" classes with respect to the Louvain's clusters. From Figure 14 it emerges very clearly that modularity-based clusters are *not* thematically uniform, since the mutual information of the two partitions is always barely greater than the mutual information of two random partitions. Thus, the apparent significance of the obtained Louvain's clusters cannot be explained by a thematic homogeneity of the clusters.

Availability of data and material The dataset used for the analysis is available at the following address https://www.cranic.it/data/supporting_material.tar.gz. Readers interested in additional information about the dataset are welcome to contact the authors.

Code availability To explore Tor we used a set of open source tools, namely tor, tinyproxy and bubing. To extract metrics and analyze data we used a set of software libraries, mainly python libraries such as igraph, numpy and scipy. To build the graphs we developed custom software in C language. Readers interested in getting our tools are welcome to contact the authors.

Declarations

Conflicts of interest The authors declare that they have no conflict of interest.

References

1. Al-Nabki, M.W., Fidalgo, E., Alegre, E., Fernández-Robles, L.: Torank: Identifying the most influential suspicious domains in the tor network. Expert Systems with Applications **123**, 212–226 (2019)
2. Al Nabki, M.W., Fidalgo, E., Alegre, E., de Paz, I.: Classifying illegal activities on tor network based on web textual contents. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pp. 35–43 (2017)
3. Alstott, J., Bullmore, E., Plenz, D.: Powerlaw: a Python package for analysis of heavy-tailed distributions. PloS One **9**(1), e85777 (2014)
4. Annessi, R., Schmiedecker, M.: Navigator: Finding faster paths to anonymity. In: IEEE European Symposium on Security and Privacy (Euro S&P). IEEE (2016)
5. Bernaschi, M., Celestini, A., Guarino, S., Lombardi, F.: Exploring and analyzing the tor hidden services graph. ACM Trans. Web **11**(4), 24:1-24:26 (2017). https://doi.org/10.1145/3008662
6. Bernaschi, M., Celestini, A., Guarino, S., Lombardi, F., Mastrostefano, E.: Spiders like onions: On the network of tor hidden services. In: The World Wide Web Conference, WWW '19, pp. 105–115. ACM, New York, NY, USA (2019). https://doi.org/10.1145/3308558.3313687
7. Biryukov, A., Pustogarov, I., Thill, F., Weinmann, R.P.: Content and popularity analysis of tor hidden services. In: Distributed Computing Systems Workshops (ICDCSW), 2014 IEEE 34th International Conference on, pp. 188–193 (2014). https://doi.org/10.1109/ICDCSW.2014.20
8. Biryukov, A., Pustogarov, I., Weinmann, R.P.: Trawling for tor hidden services: Detection, measurement, deanonymization. In: Proceedings of the 2013 IEEE Symposium on Security and Privacy, SP '13, pp. 80–94. IEEE Computer Society, Washington, DC, USA (2013). https://doi.org/10.1109/SP.2013.15
9. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment **2008**(10), P10008 (2008)
10. Boldi, P., Marino, A., Santini, M., Vigna, S.: Bubing: Massive crawling for the masses. In: Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion, pp. 227–228 (2014)
11. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J.: Graph structure in the web. Computer Networks **33**(1–6), 309–320 (2000). https://doi.org/10.1016/S1389-1286(00)00083-9
12. Celestini, A., Guarino, S.: 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, pp. 19:1–19:10. ACM, New York, NY, USA (2017). https://doi.org/10.1145/3102254.3102266
13. Christin, N.: 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, pp. 213–224. ACM, New York, NY, USA (2013). https://doi.org/10.1145/2488388.2488408
14. Clauset, A., Shalizi, C.R., Newman, M.E.: Power-law distributions in empirical data. SIAM Review **51**(4), 661–703 (2009)
15. De Domenico, M., Arenas, A.: Modeling structure and resilience of the dark network. Phys. Rev. E **95**, 022313 (2017). https://doi.org/10.1103/PhysRevE.95.022313
16. Duxbury, S.W., Haynie, D.L.: The network structure of opioid distribution on a darknet cryptomarket. Journal of Quantitative Criminology **34**(4), 921–941 (2018)
17. Franceschet, M.: Pagerank: Standing on the shoulders of giants. Commun. ACM **54**(6), 92–101 (2011). https://doi.org/10.1145/1953122.1953146
18. Ghosh, S., Das, A., Porras, P., Yegneswaran, V., Gehani, A.: 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, pp. 1793–1802. ACM, New York, NY, USA (2017). https://doi.org/10.1145/3097983.3098193
19. Girvan, M., Newman, M.E.: Community structure in social and biological networks. Proceedings of the National Academy of Sciences **99**(12), 7821–7826 (2002)
20. Griffith, V., Xu, Y., Ratti, C.: Graph theoretic properties of the darkweb. arXiv:1704.07525 (2017)
21. Guarino, S., Trino, N., Celestini, A., Chessa, A., Riotta, G.: Characterizing networks of propaganda on twitter: a case study. Applied Network Science **5**(1) (2020). https://doi.org/10.1007/s41109-020-00286-y
22. Jansen, R., Bauer, K., Hopper, N., Dingledine, R.: Methodically modeling the tor network. In: Proceedings of the 5th USENIX Conference on Cyber Security Experimentation and Test, CSET'12, pp. 8–8. USENIX Association, Berkeley, CA, USA (2012). http://dl.acm.org/citation.cfm?id=2372336.2372347

23. Khare, R., Cutting, D., Sitaker, K., Rifkin, A.: Nutch: A flexible and scalable open-source web search engine. Oregon State University **1**, 32–32 (2004)
24. Kleinberg, J., Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A.: The web as a graph: Measurements, models, and methods. In: Asano, T., Imai, H., Lee, D., Nakano, S.i., Tokuyama, T. (eds.) Computing and Combinatorics, Lecture Notes in Computer Science, vol. 1627, pp. 1–17. Springer Berlin Heidelberg (1999). https://doi.org/10.1007/3-540-48686-0_1
25. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM) **46**(5), 604–632 (1999)
26. Lehmberg, O., Meusel, R., Bizer, C.: Graph structure in the web: Aggregated by pay-level domain. In: Proceedings of the 2014 ACM Conference on Web Science, WebSci '14, pp. 119–128. ACM, New York, NY, USA (2014). https://doi.org/10.1145/2615569.2615674
27. Litvak, N., Van Der Hofstad, R.: Uncovering disassortativity in large scale-free networks. Physical Review E **87**(2), 022801 (2013)
28. Mitzenmacher, M.: A brief history of generative models for power law and lognormal distributions. Internet mathematics **1**(2), 226–251 (2004)
29. Mohr, G., Stack, M., Ranitovic, I., Avery, D., Kimpton, M.: An introduction to heritrix an open source archival quality web crawler. In: In IWAW'4, 4th International Web Archiving Workshop. Citeseer (2004)
30. Newman, M.E.J.: Mixing patterns in networks. Phys. Rev. E **67**(2), 026126 (2003). https://doi.org/10.1103/PhysRevE.67.026126
31. Norbutas, L.: Offline constraints in online drug marketplaces: An exploratory analysis of a cryptomarket trade network. International Journal of Drug Policy **56**, 92–100 (2018)
32. Owen, G., Savage, N.: Empirical analysis of tor hidden services. IET Information Security **10**(3), 113–118 (2016)
33. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Tech. rep, Stanford InfoLab (1999)
34. Sanchez-Rola, I., Balzarotti, D., Santos, I.: 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, pp. 1251–1260. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland (2017). https://doi.org/10.1145/3038912.3052657
35. Sheridan, P., Onodera, T.: A preferential attachment paradox: How preferential attachment combines with growth to produce networks with log-normal in-degree distributions. Scientific Reports **8**(1), 2811 (2018)
36. Spitters, M., Verbruggen, S., van Staalduinen, M.: Towards a comprehensive insight into the thematic organization of the tor hidden services. In: Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint, pp. 220–223 (2014). https://doi.org/10.1109/JISIC.2014.40
37. Takaaki, S., Atsuo, I.: Dark web content analysis and visualization. In: Proceedings of the ACM International Workshop on Security and Privacy Analytics, pp. 53–59. ACM (2019)
38. Vinh, N.X., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: is a correction for chance necessary? In: Proceedings of the 26th annual international conference on machine learning, pp. 1073–1080 (2009)
39. Zabihimayvan, M., Sadeghi, R., Doran, D., Allahyari, M.: A broad evaluation of the tor english content ecosystem. arXiv:1902.06680 (2019)