The Onions Have Eyes: A Comprehensive Structure and Privacy Analysis of Tor Hidden Services

Iskander Sanchez-Rola
DeustoTech,
University of Deusto
iskander.sanchez@deusto.es

Davide Balzarotti
Eurecom
davide.balzarotti@eurecom.fr

Igor Santos
DeustoTech,
University of Deusto
isantos@deusto.es

ABSTRACT

Tor is a well known and widely used darknet, known for its anonymity. However, while its protocol and relay security have already been extensively studied, to date there is no comprehensive analysis of the structure and privacy of its Web *Hidden Services*.

To fill this gap, we developed a dedicated analysis platform and used it to crawl and analyze over 1.5M URLs hosted in 7257 onion domains. For each page we analyzed its links, resources, and redirections graphs, as well as the language and category distribution. According to our experiments, Tor hidden services are organized in a sparse but highly connected graph, in which around 10% of the onions sites are completely isolated.

Our study also measures for the first time the tight connection that exists between Tor hidden services and the Surface Web. In fact, more than 20% of the onion domains we visited imported resources from the Surface Web, and links to the Surface Web are even more prevalent than to other onion domains.

Finally, we measured for the first time the prevalence and the nature of web tracking in Tor hidden services, showing that, albeit not as widespread as in the Surface Web, tracking is notably present also in the Dark Web: more than 40% of the scripts are used for this purpose, with the 70% of them being completely new tracking scripts unknown by existing anti-tracking solutions.

Keywords

privacy; dark web; browser security & privacy

1. INTRODUCTION

Informally, the *Dark Web* refers to the small portion of the *Deep Web* (the part of the Web which is normally considered to be beyond reach from current search engines) based on *darknets*. Common darknets include, among other smaller P2P networks, *FreeNet* [6], the *Invisible Internet Project* (I2P) [5], and *Tor* [2]. In the case of Tor, Tor hidden services are used to provide acccess to different applications such as chat, email, or websites, through the Tor network. In this paper, we focus in particular on the analysis of *websites* hosted on Tor hidden services — due to Tor's much larger popularity between users, which comprised around 7,000 relays or proxies by the time of this writing [4]. The Tor network is based on the onion routing technique [33] for network traffic anonymization.

Due to its hidden nature, Tor hidden services are used for a large range of (cyber)-criminals activities [13, 14, 38, 35]. Thereby, several studies [9, 27, 16, 26] focused on how to discover, access, crawl, and categorize the content of the Dark Web.

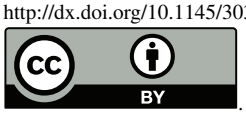
Recently, the *OnionScan* [22, 25, 24, 23] and the *DeepLight* reports [17] have analyzed some features related to the content, the size, and the connectivity of the Dark Web. While these studies have helped to better understand its nature, we still lack a complete analysis of Tor hidden services to compare their structure with the corresponding studies of the *Surface Web* [11, 29].

Similarly, while the research community has put a considerable effort to analyze the privacy and security of Tor relays [28, 12, 41, 36] and of its routing protocol [30, 18, 39, 19], a comprehensive analysis of the privacy implications at the application level and of the prevalence of fingerprinting and web tracking is still missing (although these subjects have been extensively studied for the Surface Web [32, 8, 7, 20, 21]).

To fill these gaps, in this paper we present the most comprehensive structure and privacy analysis of the Tor hidden services. Our work is divided in three parts. In the first, we present the most complete exploration of the websites hosted on the Tor hidden services performed to date. Previous measurement studies were limited just to the home pages of each site. While it is true that 80% of the websites have less then 18 URLs, according to our experiments their home pages contain only 11% of the outgoing links, 30% of the resources, 21% of the scripts, and 16% of the tracking attempts. To overcome this limitation, in our analysis we exhaustively downloaded all the reachable content for over 80% of the websites (for a total of 1.5M pages), and we completely crawled 99.46% of the sites to extract links to other domains.

In the second part of our paper, we present an analysis of the collected data looking at links and redirections, as well as at the external resources imported by onion domains

WWW 2017 Perth, Australia
ACM 978-1-4503-4913-0/17/04.
http://dx.doi.org/10.1145/3038912.305265



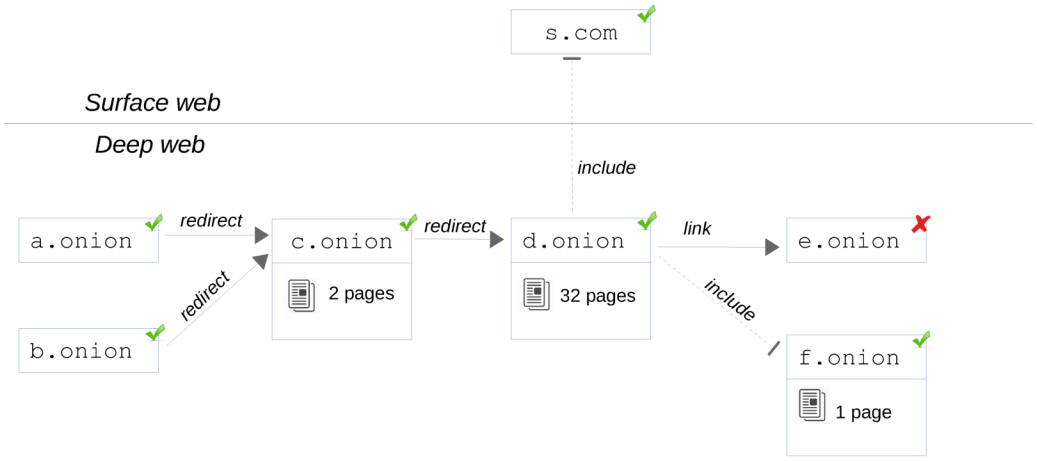


Figure 1: Tor Hidden Services Architecture Example and Clarification.

from the Tor hidden services themselves and from the Surface Web. In addition, we perform a complete structure analysis of the three connectivity graphs – links, resources, and redirections – and compare them to previous structural analyses conducted for the Surface Web.

Our experiments show that Tor hidden services are highly connected and that their internal structure is sparse, with a high number of strongly connected domains. Overall, 10% of the websites have no incoming links and a stunning 98.8% of all the discoverable domains are already included in public directories (with a single one - `tt3j2x4k5ycaa5zt.onion` pointing to over 70% of the websites we visited).

Quite surprisingly, we also discovered that Tor hidden services are more connected to the Surface Web than to other Tor hidden services. In particular, over 21% of the onion domains import resources (e.g., Javascript files) from the Surface Web. Due to these external components, we discovered that Google alone can monitor the accesses to almost 14% of the Tor hidden services in our dataset.

Since these connections can raise some privacy concerns, in the third part of the paper we analyze the privacy implications of the structure of Tor hidden services and we measure for the first time the prevalence and nature of web tracking in this environment. Using a generic web tracking analyzer, we discovered that, despite the fact that the usage of scripts in Tor hidden services is smaller than in the Surface Web, the percentage of them used for web tracking is similar. More than 75% of the onion domains that contain at least a Javascript file, perform some form of tracking. Moreover, we have found that the majority of web tracking present in Tor hidden services is not known by any anti-tracking technique.

Another interesting finding is the fact that over 30% of the tracking and fingerprinting in Tor hidden services uses scripts imported from the Surface Web. This is particularly worrying, as it may be used to follow known users when they visit anonymous websites on Tor. Finally, we discuss how the owners of the websites try to hide their tracking attempts, for instance by performing tracking in the middle of a redirection chain.

The remainder of this paper is organized as follows. Section 2 details the analysis platform and the methodology used in our Tor hidden services analysis. Section 3 describes the conducted structural analysis of the onion domains, as well as our findings. Section 4 discusses the privacy implications of our previous findings, and details the specific web tracking analysis performed in the Tor hidden services. Section 5 provides the context of this paper given the current previous work. Finally, Section 6 summarizes the main conclusions.

2. ANALYSIS PLATFORM

While the connection among different web pages is part of the nature of the Surface Web, web sites in the Dark Web are often more ephemeral and isolated among one another. This difference makes crawling the Dark Web a non-trivial task, that goes beyond simply navigating through hyper-links.

Therefore, to perform our study we manually collected a list of URLs associated to 195,748 onion domains from 25 public forums and directories. We then implemented a custom crawler to explore this seed list to collect data for our analysis and extract new domains. Our crawler can be configured to operate according to two different behaviors. In *"collection mode"* the crawler retrieves all the HTML and Javascript resources of the target domain. This mode has restrictions regarding the maximum depth and number of links it can explore for each onion domain. When these thresholds are reached, the system switches to *"connectivity mode"*, where it simply crawls the remaining pages looking for new links towards other onion domains. While in this mode, the system does not store a copy of the resources and therefore it is not restricted in its depth level (but with a maximum of 10K URLs per domain) and can visit a much larger number of pages of the target domain.

After the crawler has collected all the data, we performed an offline analysis – divided in two parts:

1. **Structure Analysis:** the goal of this analysis is to study the number of connections and their nature to better understand the overall structure and properties of the Dark Web. For instance, we measured the prevalence of links, resources, and redirections both towards

other onion domains and towards the Surface Web. We also built the complete graph for each connection type and performed a complete structure analysis, comparing the results with those obtained by analyzing the properties of the Surface Web.

2. **Privacy Analysis:** The structure analysis raised several privacy concerns, which we analyze in more details in this second analysis phase – which focus on studying the tracking ecosystem in the Dark Web. Since our goal is to obtain a general overview of web tracking in the Dark Web, rather than focusing on a specific type of tracking, we did not implement our own tracking detector but we reused instead a generic tracking analyzer capable of detecting both known and unknown tracking techniques [37].

Finally, it is important to remark that during our analysis we distinguish between *domains* and *URLs*. Figure 1 clarifies the distinction. For instance, the figure shows seven domains (`a.onion`, `b.onion`, ..., `f.onion`, and `s.com`). Domains can be hosted either on the Surface Web, or on the Dark Web (onion domains). Some of them point to website hosting actual content (e.g., `c, d`, and `f`) while other only serve to redirect users to other domains (such as `a` and `b`). For our analysis purposes, we define the *size* of a domain as the number of unique URLs served by that domain that returned HTML content. Domains can also host other resources (such as JavaScript, pictures, and CSS file) which are valid URLs but do not count as accessed URLs nor for the computation of the size.

2.1 Design and Implementation

We implemented our Dark Web crawler on top of the headless browser *PhantomJS* [1]. To prevent websites from easily identifying our crawler, we implemented a number of advanced hiding techniques already used by other systems.[1] Moreover, we cleaned cookies and the browser cache after visiting each website.

The crawler receives the main URL of an onion domain and retrieves its HTML and external scripts. To accommodate for dynamic content, the system saves the content of each page after a certain time has passed. Since the Dark Web is considerably slower than the Surface Web, instead of using a fixed waiting time, we first performed an experiment to measure the appropriate value. For this test we extracted a random sample of 25% of the initial seed domains and analyzed the loading time of their pages. Based on the results of our experiment, we configured our crawler to wait for five additional seconds after a page is completely loaded, and for a maximum of 60 seconds otherwise.

To deal with script obfuscation, we implemented a deobfuscator using *JSBeautifier*.[2] that iteratively tries to deobfuscate a page, testing at each iteration whether or not additional code has been revealed. In this way, we can also deal with multiple-layer obfuscation.

As already mentioned above, our crawler can work in two modes:

- **Collection mode:**

  In this mode the crawler collects and stores a large amount of data for further offline analysis, including all the HTTP headers, the HTML code, all scripts (both imported or embedded in the HTML), the list of performed redirections along with their origin, destination and nature (e.g., HTTP or JavaScript), and all the links (either visible or invisible) within the website. To mimic the behavior of a real user, we modified the `referrer` at each step to point to the correct origin URL.

  While in collection mode, the crawler only recursively visits ten links for each onion URL. To prioritize the links that can lead to more "interesting" pages, we first test the destination URL against a list of over 30 keywords (e.g., login, register, account, or password). Second, we dynamically compute the appearance of the link in the website according to the CSSs and other styles defined in the HTML. We use this information to rank the links based on their computed visualization size, weighted by its closeness to top of the website. Finally, the crawler in collection mode is limited to visiting internal links up to a depth of 3 layers from the homepage. In other words, the crawler collects data from the homepage, from ten of its internally linked pages, then from other ten starting from each of them, up to a depth of three steps.

- **Connectivity mode:**

  The "connectivity mode" extracts all the links (either visible or invisible) in the target website through a breadth-first exploration, without considering fragment ('`#`' position links), or files such as images or PDF documents.

  This mode is not limited by the 10-pages nor by the 3-level depth restrictions. Its major goal is to complete the crawling of a domain after the collection mode has reached its threshold. However, for practical purposes and to avoid getting stuck in endless websites that contain an infinite number of URLs (the often called *calendar effect*), we limited this mode to visit 10,000 distinct URLs for each individual domain.

2.2 Data Collection

Our collection started from an initial seed of 195,748 domains, retrieved from different sources: Tor gateways, `pastebin`, lists/directories (in the surface and dark web), `reddit` posts, and previous studies [17]. These type of sources are commonly used to discover hidden onion domains [14, 17].

Our crawler then visited each onion domain, initially running in collection mode. If this approach was unable to crawl the entire website, the crawler switches to the connectivity mode and visited the remaining internal links - this time just collecting other links without storing the content.

To gather the highest number of active domains possible, we performed our entire crawling experiment three times: twice in May 2016 (three days apart), and then again one month later in June 2016.

3. STRUCTURE ANALYSIS

The first part of our study focuses on the analysis of the structure of the Tor hidden services. To this end, we analyzed the categories and languages of the retrieved websites,

[1] `https://github.com/ikarienator/phantomjs_hide_and_seek`

[2] `http://jsbeautifier.org`

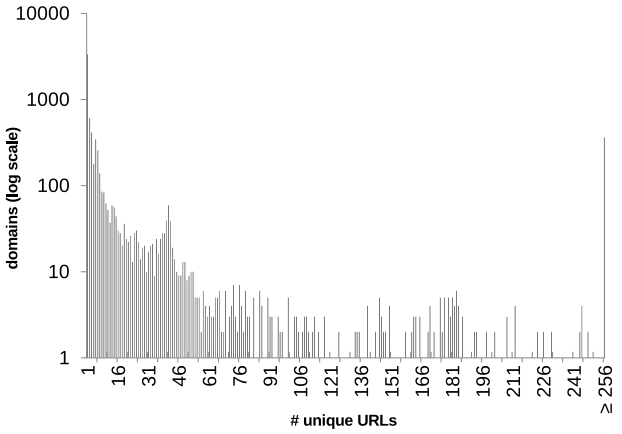


Figure 2: Distribution of URLs in each `.onion` domain.

Table 1: Most Popular Languages in Onion Domains.

Language	% Domains
English	73.28%
Russian	10.96%
German	2.33%
French	2.15%
Spanish	2.14%

Table 2: Categories in Onion Domains.

Category	% Domains
Directory/Wiki	63.49%
Default Hosting Message	10.35%
Market/Shopping	9.80%
Bitcoins/Trading	8.62%
Forum	4.72%
Online Betting	1.72%
Search Engine	1.30%

as well as their connections in terms of links, resources, and redirections. We also performed a graph analysis of the Tor hidden services and compared the results with those of similar studies performed on the Surface Web.

3.1 Size & Coverage

From our initial seed of 195,748 domains, our crawler gathered a total of 198,050 unique domains. The small difference between the two numbers confirms the nature of the dark web, where websites are mainly reached from public directories or domains posted in forums (part of our initial seed). This has two important consequences. First, that existing directories already cover 98.8% of the discoverable domains. However, Tor hidden services also include websites intended for private use, that are never linked or publicized anywhere else – and that therefore cannot be covered in our study. Second, as we explain later in this section, our experiments show that 10% of the domains have no incoming links and therefore cannot be reached by a traditional crawler. While this percentage is very large, it means that the remaining 90% of the Tor hidden services are actually interconnected, and that therefore they are not just a collection of isolated web sites.

In our three consecutive crawling attempts, we discovered that only 7,257 were active domains. This is a consequence of the short lifespan of onion websites and of the fact that the majority of onion domains collected from public sources often become unreachable after a short amount of time. Therefore, public directories contain a very large amount of outdated material, which wrongly contributes to the image of Tor hidden services.

Interestingly, 81.07% of the active domains were completely crawled in "collection mode". In this case, our system was able to retrieve and store an entire copy of the website. An additional 18.49% of the onion domains were then crawled completely in "connectivity mode" (every link and connection to external domain was successfully collected) — and only for the remaining 0.54% of the websites our system was unable to visit all URLs because they contained more than 10K internal active URLs. Overall, our crawler accessed a total number of 1,502,865 unique onion URLs corresponding to 746,298 different base URLs (i.e., without taking their parameters into account). This is the largest number of URLs visited to date in a study of Tor.

A total of 203 onion domains only performed HTTP redirections to other domains without hosting any content. For the rest, Figure 2 shows a log-scale distribution of the size of each domain. Quite surprisingly, almost half (46.07%) of them only contained a single page and over 80% contained less than 17 unique URLs. This means that vast majority of the websites in the dark web are extremely simple, and only few of them are large and complex applications, containing up to tens of thousands of different URLs.

3.2 Language & Categories

We also measured the distribution of languages and website categories within the active onion domains. Since the results are computed by analyzing each individual URL, each domain can be assigned to multiple languages and multiple categories (i.e., if a domain contains both English and Russian pages it is counted as belonging to both languages).

To obtain the actual language, we used the Google Translate API[3] and its language autodetection function. Overall, we found 63 different languages used in Tor hidden services. English was the most popular (present in 73.28% of the domains), followed by Russian, German, French, and Spanish (see Table 1). While the percentages are different, the ranking is very similar to the one of the surface web (English, Russian, German, Japanese, Spanish, French [40]) with the omission of Japanese and a larger percentage of English content. However, our results are different from the ones published in the DeepLight report [17] both in the number of languages and their distribution and ranking. This difference can be a consequence of our higher coverage, especially in the number of pages visited in each domain (for instance, a website may show only English content on the main page, but then provide also other languages by following specific links).

To identify the URL categories, we first used Google Translate to translate the page content to English, we then removed stop words[4] and used a stemming algorithm to extract the root token of similar words (e.g, work, worker, and working).

[3] `https://cloud.google.com/translate/`

[4] Using the public list at `http://www.ranks.nl/stopwords`

Table 3: Links and Resources in Onion Domains.

Links	*to Onion*	# domains linking	3,013
		# domains linked	20,621
		# domains alive linking	2,482
		# domains alive linked	6,528
	to Surface	# domains linking	2,947
		# domains linked	83,984
Resources	*from Onion*	# domains importing	466
		# domains exporting	349
	from Surface	# domains importing	1,561
		# domains exporting	2,235

Table 4: HTTP Redirection Distributions Performed in Onion Domains. Dest. means destination and D. domain.

Dest.	Type	# Source D.	# Dest. D.
Onion	HTTP	232	196
	HTML	37	22
	JS	17	12
	TOTAL	*283*	*225*
Surface	HTTP	117	124
	HTML	35	22
	JS	39	14
	TOTAL	*190*	*154*

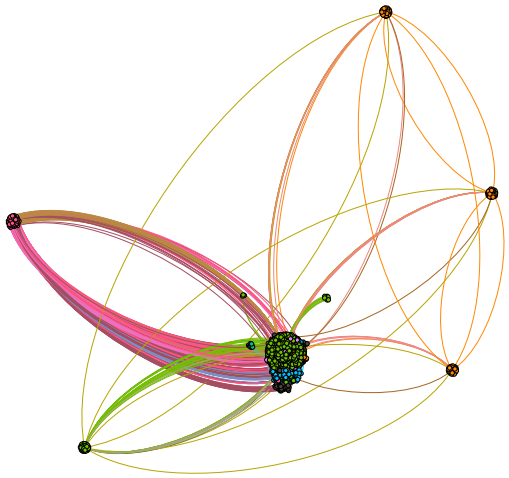


Figure 3: Links Graph of Onion Domains computed with the OpenOrd force-directed layout algorithm and colored communities through modularity. Isolated domains were removed from the figure for clearness of the representation.

Next, we modeled each URL as a *Bag of Words* [34] and performed a two-phase clustering. In the first phase, we randomly selected 10% of URLs to form the clusters by *Affinity Propagation* [15]. Then, in the second step, we computed the cluster membership of the remaining URLs. As a result, 79 different clusters were found. We manually inspected and assigned each of them to one of seven main categories (Table 2 shows their distribution). Overall, we found that 15.4% of the domains belong to more than one category. *Directory/Wiki* is the most popular category since many onion websites tend to include a resource repository of links, news, articles, videos or images. It is important to remark that directories do not necessary only link to external content, but they can also include their own. The second most popular category are websites containing a default hosting message such as "*This website is hosted by X*": this result was also observed by a previous measurement study [25].

3.3 Links, Resources, and Redirections

Table 3 shows the number, type, and prevalence of links present in onion domains, as well as the number of domains importing resources from other onion domains and from the Surface Web. Specifically, only 41.5% of the total onion domains contained HTML links to other onion domains and 40.6% contained links to the surface web. Regarding links to other onion domains, a stunning 68.34% of them were broken (i.e., the target was not reachable during our experiments) confirming again the ephemeral and disconnected nature of websites hosted in the dark web. Only 6,528 from the 7,257 active domains in our dataset were found in the onion links, indicating that over 10% of the onion domains are isolated and unreachable from other domains in our dataset. Comparing the number of links to onion and surface domains, the number of surface links was clearly higher, even considering the inactive onion links.

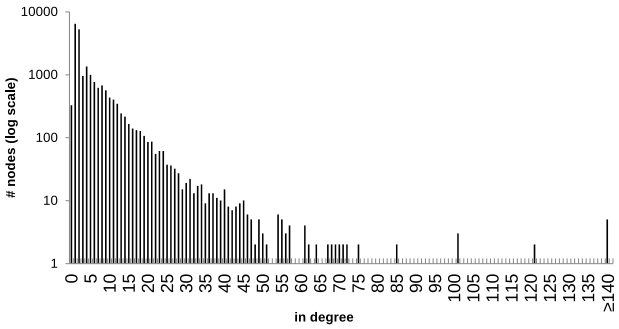
Looking at the imported resources, only 6.47% of onion domains imported resources from other onion domains, while 21.51% of them imported resources from the Surface Web. Moreover, the absolute number of unique resources imported from the Surface Web is over five times higher than the resources imported from other onion domains. As we will discuss in more details in Section 4, this can have some privacy implications on users visiting the dark web using Tor Proxies.

Another relevant finding of our study is that only 36% of the onion domains we visited contained Javascript code (either embedded in the page or standalone). Interestingly, 48% of the URLs contained no scripts at all. This shows a different picture from the Surface Web, where current statistics report that almost 94% of websites use JavaScript [40].

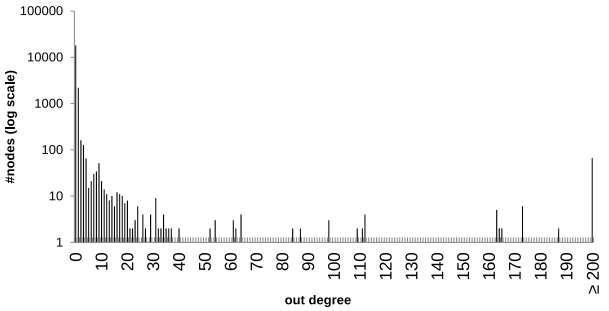
Finally, we looked at redirections. Table 4 shows that, contrary to what we observed for links and resources, redirections performed by onion domains are more likely to target other onion domains: 3.90% of the onion domains contained at least one redirection to other onions, while 2.62% redirected to the Surface Web. This second group is particularly important because these redirections can be used to de-anonymize users in certain configurations (e.g., those using a Tor proxy). Table 4 also distinguishes between HTTP and non-HTTP redirections. The majority of the sites used the `HTTP 30X` redirection method — 82% to redirect to other onion domains, and 62% to redirect to the Surface Web.

3.4 Graph Analysis

A study of the structure of the web was first conducted in 2000 by Broder et al. [11], and then revisited by Meusel et al. [29] in 2014. A first study about the connectivity of Tor hidden services has been recently presented [24], but the structure was not analyzed and only the main onion domains were taken into account by the authors. Hence, to provide a better comparison with the studies of the Surface Web, we performed a graph network analysis of the three types of connections between onion domains: links, resources, and redirections (Figure 3 shows the links graph, Figures 4a and 4b the in/out degree of the links, Figure 5a the resources

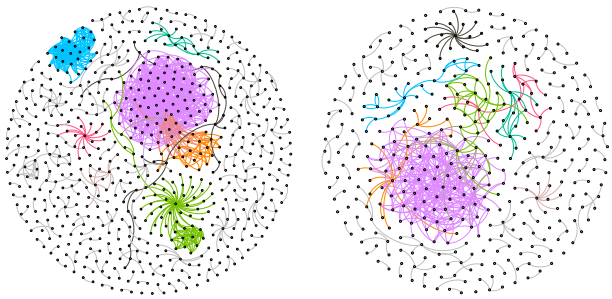


(a) Link In-Degree Distribution. The maximum in-degree has been set to 140 links or more in order to ease the understanding. Data is presented in log-scale.



(b) Link Out-Degree Distribution. The maximum in-degree has been set to 200 links or more in order to ease the understanding. Data is presented in log-scale.

Figure 4: Link In and Out Degree Distributions.



(a) Resources Graph. (b) Redirections Graph.

Figure 5: Resources and Redirection Connectivity graphs plotted using the Fruchterman-Reingold force-directed layout algorithm and colored communities through modularity.

graph, and Figure 5b shows the redirections graph), measuring the structural properties of the three resultant graphs.

The links connectivity graph contains 6,823 nodes and more than 60,000 connections. The average shortest path to traverse the entire directed network was 3.697 edges and the largest (the diameter) was 49. This relation implies that it is highly connected but disperse, as we can also deduce from the high average connection degree and its low density (near zero). The graph has a high connectivity, containing 88.03% of strongly connected nodes. Indeed, the number of communities (10), its modularity (0.293), and clustering coefficient (0.481) are relatively small, while the centrality is high (1.061). In conclusion, regarding links, the analysis of their network clearly indicates that the highly connected graph is due to a few onion domains with high in/out degree (as seen in Figures 4a and 4b). The in-degree power law of the links within the dark web is 8.588, whereas the out-degree power law is outside the 2-to-3 *scale-free networks* range — but close to it (3.234).

By comparing our work with the Surface Web structure described in previous studies [11, 29], we can observe few notable differences. First, onion domains are less prone to be linked by others, while their out-degree is smaller but similar to the one in the surface. These stats are aligned with the assumption about the undercover nature of onion domains, which explains the smaller "in" degree. We also tested whether the connections between the Tor hidden services follows the *bow-tie structure* characteristic of the Surface Web. This structure is composed of strong connected components that are linked but not accessed by an *in* cluster, and linked to the *out* cluster. Our results indicate, that even though there is a high number of strongly connected domains, there is an absence of clearly defined *in* or *out* clusters.

Next, we looked at the important *hubs* that are connected to a large part of the Tor hidden services graph. Two main hubs have a high number of incoming connections, and they are linked by a 13.65% and 5.29% of the nodes, respectively. The first one was a hosting service while the second was a bitcoin blockchain explorer/wallet. In the case of outgoing links, 8 onion domains linked to more than 50% of the network each domain. In particular, a single directory domain linked to 71.16% of the nodes.

Finally, resources and redirections graphs (shown in Figure 5) present similar structural values. The resources graph was composed of 664 nodes and 1294 edges, whereas the redirections graph contained 416 nodes and 554 edges. The out-degree power law was 4.951 in the case of resources and 5.094 in the case of redirections, while the in-degree was 4.838 for resources and 5.062 for redirections. The average shortest path for resources was smaller (2.715) than in redirections (3.596), and both of them lower than the one of the links. The diameter was smaller in this case (7 for resources and 10 for redirections).

These lower values are due to the size of these two graphs being much smaller than the links connectivity graph. However, both of them are still highly connected: 82.83% of the nodes are strongly connected in resources and 84.88% in redirections. Networks are not as sparse as in the case of links, the clustering coefficient (0.060 in the case of resources and 0.013 in the case of redirections), and the number of communities are high (156 in the case of resources and 97 in the case of redirections), while the network centrality is small (0.038 in the case of resources and 0.017 in redirections). In these cases, as we will discuss in Section 4 there are serious privacy implications due to the high connectivity of onion domains. Regarding connections amid communi-

ties, resources have 28 inter-community edges while only 6 redirections. In both cases, most of these connections are between the famous onion search engine `Grams` and its bitcoin cleaner `Helix` with directories and markets.

4. PRIVACY ANALYSIS

In the second part of our study we look at privacy implications related to the structure of the Tor hidden services and measure the web tracking prevalence and its nature within Tor hidden services. In order to measure how common web tracking is in the Dark Web as well as finding out its nature, we used the web tracking analyzer proposed by Sanchez-Rola & Santos [37] to analyze all the scripts retrieved by our crawler.

4.1 Dark-to-Surface Information Leakage

One of the main surprises of our structural analysis experiments is the tight connection that exists between Tor hidden services and the Surface Web. As we discussed in Section 3, a large number of websites contain links, import external resources, and even redirect to websites outside the onion network. While this may appear innocuous, it has important consequences for the cover nature of these websites.

Users usually visit Tor hidden services either by using a browser connected to the Tor network or by using one of the several Tor proxies (e.g., Tor2Web [3]). These services, available on the Surface Web, act as gateways to bridge incoming connections towards the Tor network. They are popular because they do not require any specific set up or installation of additional software.

When a user is using one of these proxies, instead of typing in the address bar the onion URL (e.g., `duskgytldkxiuqc6.onion`), she replaces its domain with a different one that points to the proxy service (such as `.onion.to`, `.onion.city`, `.onion.cab`, or `.onion.direct`). The proxy then acts as an intermediary to fetch the resource from the Dark Web and forward it to the user. In addition, it transparently rewrites any onion-related URL in the returned page, to append one of the aforementioned proxied domains.

The fact that Tor proxies do not provide the same privacy guarantees of a real connection to the Tor network is well known, and even advertised on their sites. However, the main privacy issue discussed so far was that the proxy knows the user's IP address and can monitor its traffic. Hence, the user needs to trust the proxy. However, our measurement shows that from a privacy perspective, there is also another important consequence of using these proxies. In fact, in many cases not just the proxy knows the IP address of the user, but even third-parties and the target websites can get access to this information.

4.1.1 Links, Resources, and Redirections

When an onion website imports a resource from the Surface Web, its URL is not rewritten by the Tor proxy and therefore it is fetched by the user browser in the usual way, bypassing the anonymization network. This has two important consequences. First, since resources are loaded each time the website is visited, the destination domain can monitor the traffic of the source onion domain. Despite the fact that this problem is already known, we have, for the first time, measure its impact. According to the data we collected, a handful of popular surface domains can monitor a remarkable percentage of the traffic towards onion domains. Google (13.20%), Facebook (1.03%), and Twitter (0.88%) alone cover 13.39% of the onion domains. Moreover, while these statistics are anonymous if the user is connected through the Tor network, they are not when proxies are used to access Tor hidden services. In other words, Google alone can monitor the IP address of any client who visits over 13% of the onion websites, if the user uses a Tor proxy.

The second consequence is that any onion website can simply import a library or an image from the surface web to transparently get access to every user's IP address visiting the domain using a Tor proxy. Overall, since over 21% of the onion websites import resources from the surface Web, we believe that this is a very widespread issue that should further motivate users not to use Tor proxies. Redirections from onion domains to the Surface Web are less common, but still account for a relevant percentage of the websites (2.6%). In this case, if the user is using a Tor proxy, she is redirected to the surface, thus losing completely her anonymity.

4.1.2 Countermeasures

The risk of using Tor proxies is well known, but our study shows that it is even more severe than we previously thought. For users, the obvious solution is to avoid using Tor proxies and connect directly to the Tor network. Otherwise, they need to be aware that their identity can be tracked by the proxy, the target website, or by several large companies such as Google and Facebook.

Website-related privacy issues found by our structure analysis are due to the ability of a destination domain (the target of a link or imported resource) to know from which domain the request is coming and, therefore, reveal its existence to others. In order to avoid this website disclosure there are several options. First, the developer can avoid having any external resources, links, or redirections to make impossible for an external domain to know about its existence or to monitor its traffic. If the website needs to use external resources, the developers should copy them in their website, checking that these resources do not fetch additional components from other hosts. Second, it is possible to maintain the external connections but hide the `http_referrer` HTTP header. For example, the attribute `no-referrer` can be used to hide this property when fetching additional resources or following a link. Surprisingly, only 0.54% of the onion domains used this technique to protect their origin when importing external resources.

4.2 Tracking

Tracking Prevalence

As mentioned in Section 2, we used a previously proposed tracker analysis tool [37] to analyze the scripts we retrieved from the onion websites. Using this tool, we compute the tracking prevalence with regards to every script (Table 5) and every onion domain (Table 6). The tool we use models the source code as a Vector Space Model and then it computes the cosine similarity amid the inspected script and known tracking scripts within a database. If there are no matches, it uses a machine-learning approach to categorize unknown tracking scripts.

By using this tool we can divide the tracking scripts in three categories: (i) *Known* scripts, which are at least 85% similar to already known and blacklisted tracking scripts;

Table 5: Prevalence of Web Tracking Scripts in Scripts.

Type	# Scripts	% of All Scripts	Unique
Tracking	*118,675*	*44.02%*	*12,285*
– Known	22,736	8.43%	469
– Unknown Blacklisted	12,816	4.76%	1,392
– Completely Unknown	83,123	30.83%	10,438
Non tracking	*150,917*	*55.98%*	*13,053*
TOTAL	*269,592*	*100.00%*	*25,338*

Table 6: Prevalence of Web Tracking in Onion Domains.

Type	# Domains	% Domains with Scripts	% All Domains
Tracking	*1,992*	*76.82%*	*27.49%*
- Known	501	19.32%	6.92%
- Unknown Blacklisted	436	16.81%	6.02%
- Completely Unknown	1,886	72.73%	26.03%
Non tracking	*1,886*	*71.96%*	*25.76%*
No scripts	*4,652*	*N/A*	*64.21%*

(ii) *Unknown Blacklisted* scripts, which were not previously known but they were imported from blacklisted tracking domains; and (iii) *Completely Unknown* scripts that were not previously known nor imported from a blacklisted domain.

According to Table 5, nearly half (44.02%) of the scripts present in onion domains performed diverse types of web tracking such as analytics, cookies, or device fingerprinting. This ratio is similar to what has been previously reported for the prevalence of Surface Web tracking [37].

More than 75% of the onion domains that contain at least one script, perform tracking (see Table 6). However, considering that 64.21% of the onion domains with HTML did not use any script, only 27.49% of all onion domains used tracking. The prevalence of unknown tracking candidates in websites was also the highest: 94.68% of tracking onion domains used at least one unknown script, while known scripts appeared in 25% of the domains, and scripts from blacklisted domains in 21%.

Tracking Specifics

We collected 118,675 tracking scripts but, as shown by the last column in Table 5, only 10% of them were unique. We also checked where the tracking scripts were hosted: as a script file in the onion domain, embedded in the HTML, or in third-party domains. The majority of them was hosted in the onion domain itself: 40.39% as a separate resource and 26.11% embedded in the HTML.

Finally, to understand the tracking scripts, we performed a cluster analysis with the same methodology used in Section 3. In this case, we started by clustering 957 known tracking scripts using the *Affinity Propagation* algorithm [15]. Then, we computed the closest cluster for each of the tracking scripts found in our dataset. By manually analyzing the most prevalent clusters from the resulting 106 clusters, we found that 17.10% of the scripts performed statistics, 15.04% performed stateless tracking, 10.48% were used for targeted advertisement, 10.08% for web analytics, and 7.22% were stateful tracking scripts.

Hiding Techniques

We then analyzed the use of different hiding techniques, including (i) obfuscation, (ii) embedding the script into the HTML, and (iii) placing web tracking scripts in the middle of a redirection chain (i.e., in a HTML resource which is neither the source nor the origin of a multi-step redirection).

Table 7: Surface Third Party Tracking Scripts Prevalence. The percentage of scripts is computed with regards to the total number tracking scripts coming from the surface web.

Type	# Scripts	% Scripts
Surface Known	14,990	38.86%
Surface Unknown Blacklisted	12,816	33.22%
Surface Completely Unknown	10,769	27.92%
Total Surface Tracking	*38,575*	*100.00%*

As we already discussed, our crawler uses a de-obfuscator to process each collected script file. However, to our surprise, we discovered that only a 0.61% of the tracking scripts were obfuscated.

Script embedding, which consists in copying the source code of a tracking script and embedding it as `<script>` in the HTML, is a common anti-tracking solutions to bypass URL-based detection schemes. In our dataset, 16.28% of the samples were embedded in the HTML. Among these samples, there are well-known web tracking scripts such as `dota.js` or `analytics.js`. For example, `dota.js`, known for performing canvas fingerprinting [7], was always embedded in the HTML. In comparison, Google's `analytics.js` was instead embedded in only 0.66% of the samples.

Finally, we observed an interesting technique in which the tracking script was hidden in intermediate URLs part of a multi-step redirection chain. For instance, a page `A` can redirect the user to `B` – which performs the tracking and then redirects again the user to a third page `C`. This setup evades those systems that load a URL and only perform the analysis on the final resources. While this technique was not widely used, a significant 1.67% of the scripts (280 unique) were found to be hosted in intermediate HTMLs.

Surface Third-party Web Tracking

Table 7 shows the number of scripts and its distribution between known, unknown blacklisted, and completely unknown tracking scripts. Web tracking scripts coming from the surface web represented the 32.50% of all the web track-

ing present in the dark web and 97.04% of all the third-party tracking. Obviously, every script from a blacklisted domain was loaded from the surface web. However, the number of known web tracking imported from the surface domains is particularly high: 65% of all the already known scripts.

This is a serious issue, as adopting web tracking techniques from the Surface Web may be used to de-anonymize users. For instance, if a Tor hidden services uses the same tracking script of a site on the Surface Web, then the script can fingerprint the user even if she connects through Tor and then identify her when she connects to other websites on the Surface Web. In this case, it is important to use a browser hardened against fingerprinting, such as the Tor browser.

The vast majority of the tracking scripts imported from the Surface Web were from Google (43%) followed by Facebook (3.2%) and Twitter (1.9%). In total, we counted 146 unique surface domains. As we already discussed for imported resources, these surface domains may monitor traffic from an important number of Tor hidden services.

5. RELATED WORK

Relay security [28, 12, 41, 36] and traffic analysis [31, 30, 19, 39] have been very popular lines of work regarding the security and anonymity of the Tor network. Our work focuses instead on the analysis of websites hosted in the Tor network (the so-called Dark Web).

One of the very first works that studied the nature of the Dark Web was presented by Bergman [9]. In his work, the author introduced and analyzed for the first time different characteristics of the Tor hidden services, including size, content, or their ability to remain covert.

Cyber-criminal activities in Tor hidden services had been analyzed in two recent reports by Ciancaglini et al. [13, 14], showing that this venue is commonly used to perform illicit activities by different types of criminals. Moreover, in their second report, the authors measured the distribution of several common features of the Dark Web, such as languages, market products, and criminal categories. Soska & Christin [38] presented a long-term analysis of anonymous marketplaces, providing a comprehensive understanding of their nature and their evolution over time. In a similar vein, Lewis [22, 25, 24, 23] and Intelliagg & Darksum [17] performed a number of preliminary studies of the typology of these networks and its privacy issues.

A crawling methodology for isolated Tor hidden services was also presented by Biryukov et al. [10]. But this methodology required to monitor the exit nodes of the Tor darknet, and therefore we preferred to use a less invasive approach in our work.

None of the aforementioned studies performed a complete structure analysis of the Dark Web, and neither they analyzed the privacy implications or the web tracking activity performed by onion websites. In addition, these studies had a much limited coverage of the nature of Tor hidden services, due to the fact that they draw their conclusions by accessing only the homepage of the different onion domains.

6. CONCLUSIONS

In this paper we presented the first structure and privacy analysis of Tor hidden services — based on the largest experiments performed to date in this environment. We analyzed the prevalence of languages and categories, and the structure of the resultant Tor hidden services connection graph. We found that the Dark Web is highly connected but it does not exhibit the *scale-free network* and *bow-tie structure* of the Surface Web.

Connections to the Surface Web from onion domains not only exist, but they are extremely common, even more than towards other onion domains. In addition, more than 20% of the onion domains imported resources from the Surface Web. In the paper we also measure the impressive prevalence of web tracking, as nearly half of the scripts (70% of which completely unknown) were tracking and were used by nearly 30% of the onion domains.

Acknowledgments

This work is partially supported by the Basque Government under a pre-doctoral grant given to Iskander Sanchez-Rola.

7. REFERENCES

[1] PhantomJS. `http://phantomjs.org/`.

[2] Tor Project: Anonymity Online. `https://www.torproject.org`.

[3] Tor2web: Browse the Tor Onion Services. `https://www.tor2web.org/`.

[4] TorMETRICS. `https://metrics.torproject.org`.

[5] I2P: The Invisible Internet Project. `https://geti2p.net/`, 2016.

[6] FreeNet. `https://freenetproject.org`, Accessed: September 2016.

[7] G. Acar, C. Eubank, S. Englehardt, M. Juarez, A. Narayanan, and C. Diaz. The web never forgets: Persistent tracking mechanisms in the wild. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2014.

[8] G. Acar, M. Juarez, N. Nikiforakis, C. Diaz, S. Gürses, F. Piessens, and B. Preneel. FPDetective: dusting the web for fingerprinters. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2013.

[9] M. K. Bergman. White paper: the deep web: surfacing hidden value. *Journal of electronic publishing*, 7(1), 2001.

[10] A. Biryukov, I. Pustogarov, and R.-P. Weinmann. Trawling for tor hidden services: Detection, measurement, deanonymization. In *IEEE Symposium on Security and Privacy (Oakland)*, 2013.

[11] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Computer networks*, 33(1):309–320, 2000.

[12] S. Chakravarty, G. Portokalidis, M. Polychronakis, and A. D. Keromytis. Detecting traffic snooping in tor using decoys. In *Proceedings of the International Workshop on Recent Advances in Intrusion Detection (RAID)*, 2011.

[13] V. Ciancaglini, M. Balduzzi, M. Goncharov, and R. McArdle. Deepweb and Cybercrime: It's Not All About TOR. `http://www.trendmicro.com/cloud-content/us/pdfs/security-intelligence/white-papers/wp-cybercrime-and-the-deep-web.pdf`, 2013.

[14] V. Ciancaglini, M. Balduzzi, R. McArdle, and M. Rösler. Below the surface: Exploring the deep web. https://www.trendmicro.com/cloud-content/us/pdfs/security-intelligence/white-papers/wp_below_the_surface.pdf, 2015.

[15] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.

[16] B. He, M. Patel, Z. Zhang, and K. C.-C. Chang. Accessing the deep web. *Communications of the ACM*, 50(5):94–101, 2007.

[17] Intelliagg and Darksum. DEEPLIGHT: shining a light on the dark web. http://www.deep-light.net/, 2016.

[18] R. Jansen, F. Tschorsch, A. Johnson, and B. Scheuermann. Anonymously deanonymizing and disabling the tor network. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2014.

[19] A. Kwon, M. AlSabah, D. Lazar, M. Dacier, and S. Devadas. Circuit fingerprinting attacks: Passive deanonymization of tor hidden services. In *Proceedings of the USENIX Security Symposium (SEC)*, 2015.

[20] P. Laperdrix, W. Rudametkin, and B. Baudry. Beauty and the beast: Diverting modern web browsers to build unique browser fingerprints. In *Proceedings of the IEEE Symposium on Security and Privacy (Oakland)*, 2016.

[21] A. Lerner, A. K. Simpson, T. Kohno, and F. Roesner. Internet jones and the raiders of the lost trackers: An archaeological study of web tracking from 1996 to 2016. In *Proceedings of the USENIX Security Symposium (SEC)*, 2016.

[22] S. J. Lewis. Onionscan report: April 2016. https://onionscan.org/reports/april2016.html/, 2016.

[23] S. J. Lewis. Onionscan report: July 2016 - https somewhere sometimes. https://mascherari.press/onionscan-report-july-2016-https-somewhere-sometimes/, 2016.

[24] S. J. Lewis. Onionscan report: June 2016. https://mascherari.press/onionscan-report-june-2016/, 2016.

[25] S. J. Lewis. Onionscan report: May 2016. https://onionscan.org/reports/may2016.html, 2016.

[26] W. Liu, X. Meng, and W. Meng. Vide: A vision-based approach for deep web data extraction. *IEEE Transactions on Knowledge and Data Engineering*, 22(3):447–460, 2010.

[27] J. Madhavan, D. Ko, Ł. Kot, V. Ganapathy, A. Rasmussen, and A. Halevy. Google's deep web crawl. *Proceedings of the VLDB Endowment*, 1(2):1241–1252, 2008.

[28] D. McCoy, K. Bauer, D. Grunwald, T. Kohno, and D. Sicker. Shining light in dark places: Understanding the tor network. In *Proceedings of the Privacy Enhancing Technologies Symposium (PETS)*, 2008.

[29] R. Meusel, S. Vigna, O. Lehmberg, and C. Bizer. Graph structure in the web-revisited. In *Proceedings of the International World Wide Web Conference (WWW)*, 2014.

[30] P. Mittal, A. Khurshid, J. Juen, M. Caesar, and N. Borisov. Stealthy traffic analysis of low-latency anonymous communication using throughput fingerprinting. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2011.

[31] S. J. Murdoch and G. Danezis. Low-cost traffic analysis of tor. In *Proceedings of the IEEE Symposium on Security and Privacy (Oakland)*, 2005.

[32] N. Nikiforakis, A. Kapravelos, W. Joosen, C. Kruegel, F. Piessens, and G. Vigna. Cookieless monster: Exploring the ecosystem of web-based device fingerprinting. In *Proceedings of the IEEE Symposium on Security and Privacy (Oakland)*, 2013.

[33] M. G. Reed, P. F. Syverson, and D. M. Goldschlag. Anonymous connections and onion routing. *IEEE Journal on Selected areas in Communications*, 16(4):482–494, 1998.

[34] G. Salton, A. Wong, and C.-S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[35] A. Sanatinia and G. Noubir. Onionbots: Subverting privacy infrastructure for cyber attacks. In *Proceedings of the IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, 2015.

[36] A. Sanatinia and G. Noubir. Honey onions: a framework for characterizing and identifying misbehaving tor hsdirs. In *Proceedings of IEEE Conference on Communication Networks Security*, 2016.

[37] I. Sanchez-Rola and I. Santos. Known and Unknown Generic Web Tracking Analyzer: A 1 Million Website Study. Technical report, DeustoTech, University of Deusto, 2016.

[38] K. Soska and N. Christin. Measuring the longitudinal evolution of the online anonymous marketplace ecosystem. In *Proceedings of USENIX Security Symposium (SEC)*, 2015.

[39] Y. Sun, A. Edmundson, L. Vanbever, O. Li, J. Rexford, M. Chiang, and P. Mittal. RAPTOR: routing attacks on privacy in Tor. In *Proceedings of the USENIX Security Symposium (SEC)*, 2015.

[40] W. T. Surveys. Usage of javascript websites. https://w3techs.com/technologies/.

[41] P. Winter, R. Köwer, M. Mulazzani, M. Huber, S. Schrittwieser, S. Lindskog, and E. Weippl. Spoiled onions: Exposing malicious tor exit relays. In *Proceedings of the Privacy Enhancing Technologies Symposium (PETS)*, 2014.