A Synopsis of Critical Aspects for Darknet Research

Florian Platzer
florian.platzer@sit.fraunhofer.de
Fraunhofer SIT
Darmstadt, Germany
ATHENE
Darmstadt, Germany

Alexandra Lux
alexandra.lux@sit.fraunhofer.de
TU Darmstadt
Darmstadt, Germany
University of Hohenheim
Hohenheim, Germany

ABSTRACT

Descriptives of *the darknet*, and in particular of the Tor network, appear inconsistent and implausible in nature. In order to gain insight into how these conflicting results are produced, the goal of this study is to review previous research on the matter with regard to terminology used, methodology of sample collection and the analysis of the data. Our results indicate six critical aspects that in particular pertain to (A) an inconsistent use of terminology, (B) the methodology with which the sample was gathered, as well as the handling of (C) short-lived services, (D) botnet command and control servers, (E) web services with undetermined content and (F) duplicates of onion services. Further, we include a small case study on darknet marketplaces to demonstrate how reports concerning the number of a certain category can easily mislead. Through the implications of these aspects the presented description of Tor does not necessarily reflect the actual nature of Tor.

CCS CONCEPTS

• **Networks** → *Peer-to-peer networks.*

KEYWORDS

Darknet, Dark web, Tor, Review

ACM Reference Format:
Florian Platzer and Alexandra Lux. 2022. A Synopsis of Critical Aspects for Darknet Research. In *The 17th International Conference on Availability, Reliability and Security (ARES 2022), August 23–26, 2022, Vienna, Austria.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3538969.3544444

1 INTRODUCTION

Coverage of *the darknet*, whether in the media or even in academic studies, often focuses on aspects of crime. While neutral descriptions of anonymization technologies or privacy enhancing technologies such as Tor also emphasize the dual use character, this aspect is often neglected when it comes to media reports or scientific studies. Media reports are determined by their news value and thus focus on particular newsworthy events such as the take down of a drug marketplace. Scientific studies concerning darknet technologies such as Tor are often subject to the narrative of wanting to *shine light into the darkness*, the anonymous part of the Internet, *the darknet*. In order to do so and provide descriptives, a common first step is to pose research questions, such as the size and content of the darknet. However, results show an inconsistent picture. While some results are conflicting, others appear implausible.

To this end, the goal of this study is to identify aspects that are of particular importance when conducting research in the context of *the darknet*. Specifically, we are interested in research concerning Tor.

In order to illustrate these aspects and demonstrate how they were treated in previous research, we identified central studies that conducted research on the matter. We approach this problem by focusing on the general logic of a research process and thus define the following three references with respect to the research process: (1) the corresponding terminology, (2) the methodology to collect the sample, and (3) the analysis of the data. We identify critical aspects in these stages and finally discuss the implications with regard to (a) the subsequent research process and (b) results and interpretations of the works.

In the following section, we will hence explicate the technical background, relevant terminology of darknet technologies, as well as methods of how to collect onion addresses. Subsequently, in Section 3, we will depict the status quo of research concerning descriptives (e.g. size and content) of the Tor network. Following, in Section 4, we will discuss critical aspects of research in the context of the Tor network that are prone to error and potentially result in a misrepresentation of the Tor network. To exemplify this, in Section 5, we include a small case study to show how reports about the amount of marketplaces to be found on the Tor network can easily be misleading. Finally, we close with a discussion of the implications of the previously named critical aspects of research in the context of the Tor network and include suggestions of how to prevent these errors in further work in Section 6.

2 BACKGROUND

Darknets are overlay networks that build on the infrastructure of the Internet. Example technologies for a darknet are I2P, Freenet or Tor. Within these networks, Internet services or the sharing of files can be provided anonymously. Websites in the I2P network are called *eepSites* and in Freenet *Freesite*. Services offered in the Tor network are called *onion services* (formerly, hidden services) and, in addition to web services, can also offer services such as SSH, FTP, email or IRC chats. Each service is accessible under a specific port number. These can vary, but there are standardized default port numbers for each type of service. For example, FTP services are accessible under port number 20 and 21, and web services are accessible under port number 80 or 443. The part of the darknet that offers web services, i.e. HTTP(S) services are called *dark web.*

ARES 2022, August 23–26, 2022, Vienna, Austria
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9670-7/22/08.
https://doi.org/10.1145/3538969.3544444

The part of the Internet that does not belong to the darknet is called *clearnet*.

The best-known and most widespread anonymization network is Tor, which is the focus of this study exclusively. In the Tor network, there are two versions of URL addresses (onion addresses) of onion services. The old v2 address is 16 characters long. The new v3 address is 56 characters long. Both versions have 32 possible values per character. The v2 address has been deprecated since the end of 2021 and is no longer supported on the Tor network. In the following subsection we describe methods how to collect onion addresses.

2.1 Collecting of onion addresses

In the literature there are two methods to collect onion addresses in the Tor network.

The most common variant for collecting onion addresses is by applying a web crawler. A web crawler is a program that automatically scans a web page, searches for links (URLs or onion addresses) and retrieves them in turn. Web crawlers can search and analyze the content of a web page according to certain criteria or download the entire web page. As a starting point, a web crawler is given one or more web pages as seed. These web pages are then searched for other onion addresses. By the recursive call of the web pages, further onion addresses can be found and collected depending on the settings and programming properties. It is important to note that only addresses published on web pages can be found. Thus, addresses that are not published on web pages will not be found by the web crawling.

The second method to collect onion addresses in the Tor network is to actively set up Tor nodes. Each onion service publishes information about the accessibility of the service as *descriptors* to so-called directory servers. These servers are arranged in a Distributed Hash Table (DHT). Each Tor node can serve as a directory server for multiple onion services for a certain period of time. For this, a Tor node must be assigned the HSDir flag. The assignment depends on the properties of the Tor node (e.g. available bandwidth or uptime of the node) [22]. Tor onion services sign the published descriptor with a key which is represented by their onion address. For this reason it is possible to collect the onion addresses of the onion services. Thus, by setting up several Tor nodes, it is possible to analyze multiple published descriptors and consequently extract the respective onion addresses. This procedure allows to collect currently available onion addresses in the Tor network. However, since the final transition from v2 to v3 addresses of onion services in 2021, the descriptors are signed with a blind signature [19]. Thus, collecting onion addresses through directory server is no longer possible. In the following, we use the term *DHT* when we refer to the collection of onion addresses via directory servers.

3 RELATED WORK

There are various studies about the descriptive characteristics of the Tor network in the past. Various studies crawled all onion addresses that can be found on Tor web pages or collect onion addresses through DHT. Based on this, studies analyzed the total size of the Tor darknet [18, 19] or the dark web [4, 14], categorized the content of the websites [1, 2, 5, 6, 8, 11–14, 17, 18, 23, 26], examined the links of the individual websites among each other [2–5, 7, 23, 26] or investigated the popularity of onion services [8, 9, 18, 25]. Other aspects of research included language, protocols or services, operating systems or service runtimes. Most of these works were able to identify a quite higher number of onion services than those works, that analyzed all found onion addresses for duplicates or other coherencies between the onion services [2, 6, 10, 19, 25, 27]. These results show that the actual number of onion services must be significantly smaller than the analyses that do not take this into account. Figure 1 shows the amount of all v2 onion addresses that Tor Metrics estimated [21]. Tor Metrics extrapolates network totals with only 1% of all relays in the Tor network by using the DHT approach [15]. The number of unique v2 onion addresses for a single day ranged from 25,000 in 2013 over 100,000 in 2017 to 175,000 in 2021.

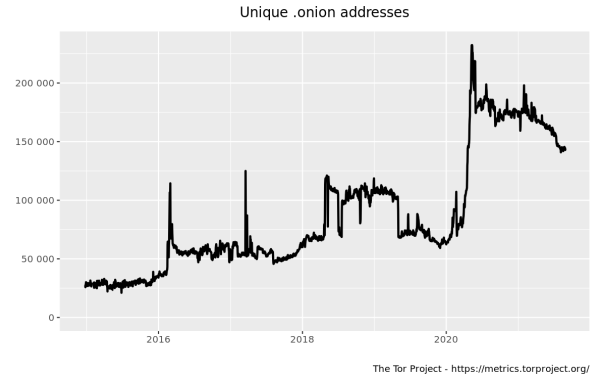

Figure 1: Amount of v2 onion addresses estimated from Tor Metrics

Owenson et al. pointed out that the Tor darknet is about half the size of the absolute number of onion services that exist [19]. Some statistics present the sum of all onion services that could be observed on one day. However, they showed that more than half of all services disappeared again within this day. Many of them were never accessible.

In 2014, Owen and Savage observed over 80,000 onion services within a period of 6 month [18]. They estimated the number of onion services to be 45,000 at any point of time. Further, fewer than 10,000 onion services are persisted with a lifetime over 18 months. Steinebach et al. also discovered that the size of the Tor network is quite smaller than the actual amount of unique onion addresses show [25]. They collected 173,190 unique onion addresses in 2018 and estimated that only half of all services are permanent services, i.e. services that run continuously longer than two month.

Even those who have analyzed websites exclusively, and therefore the dark web, recognize that there are fewer websites than absolute numbers would suggest.

Aoki and Goto crawled the dark web for a time of over two and a half years from June 2018 to January 2021 [4]. They estimated the size of the dark web within 32 measurements. The values are from 14,509 to 96,034 onion addresses with an average size of 40,848 onion addresses.

However, this is not the number of different websites that a Tor user can actually browse. In 2020, Barr-Smith and Wright investigated the amount of phishing web services within the Tor network [6]. They analyzed 11,533 onion addresses of which over 51% were imitations of other websites.

Brenner et al. tried to figure out how individual websites are associated with each other [10]. They analyzed 258 single vendor shops crawled in 2020 and calculated a similarity of all onion services in order to figure out if multiple single vendor shops belong together. They showed that only 49% are individuals. 31% are assigned to seven similarity groups [20] and 20% of all analyzed vendor sites are duplicates. Brenner et al. showed that the number of single vendor shops are quite high, but the actual supply is much smaller.

That the number of websites of individual categories of website content in the Tor network is quite high is shown by other works. Many researchers are trying to find out what content exists on the dark web and how it is distributed [1, 2, 5, 8, 11, 12, 14, 17, 18]. All of them collected onion addresses by crawling the dark web or through the DHT approach. They used several methods to identify the content category of the crawled websites. These studies show the distribution of absolute onion services but not the actual supply since none of these studies first examined the onion services for e.g. duplicates.

Table 1 compares previous work that collected and evaluated onion addresses. Indicated are the years of work performed as well as the years of publication of the studies. The column *Method* indicates which method was used to collect onion addresses. Furthermore, the table indicates how many onion addresses were found (*addresses found*), how many of them were online (*online*), and how many of them could be analyzed (*analyzed*). If the contents of Tor websites were categorized, the number of analyzed onion services is given at *Content categorized.* Works that investigated how many onion addresses were long-lived services have entries at *Permanent services.* Parentheses specify the time at which an onion service is considered as a long-lived service. Statements about the size of the Tor darknet are in the column *size.* A percentage indicates the actual size of the Tor darknet estimated by the authors compared to the total number of existing onion addresses. *[DW]* denotes that only web services were analyzed and therefore the size of the dark web is indicated. If no information about these values is available in the paper, this is marked with *n/a.* A dash (-) indicates that this item has not been considered in the work. Values with a star (*) indicate that these values have not been explicitly reported in the papers, but could be read or calculated by given statements, tables or graphs.

4 CRITICAL ASPECTS FOR DARKNET RESEARCH

There are several aspects that make it difficult to (1) prepare a representative set of onion services, (2) evaluate them in a sensible manner, and (3) conclude with a representative statement of the analysis. Below we have collected various critical aspects based on a set of previous research papers.

4.1 Inconsistent use of terminology

In the literature, terms such as *darknet*, *dark web* or *deep web* are not used consistently. The *darknet* describes the infrastructure that spans the anonymity network. Within this network there are servers that provide various services. The *dark web* is the part of the darknet that includes all web services (darknet websites). The relationship between the darknet and the dark web is comparable to the Internet and the World Wide Web. The *deep web* describes the part of the World Wide Web that cannot be found by standardized web search engines such as Google or Bing.

Several studies do not distinguish between darknet and dark web and equate both terms [1, 2, 16]. Often, the dark web is described as a part of the deep web [1, 2, 7, 16, 23, 26], or no distinction is made between deep web and dark web [13]. Others use the term darknet exclusively, but analyze the dark web [5, 6]. This inconsistent terminology makes it difficult to compare the results of these studies. Furthermore, results are interpreted in inaccurate contexts, thus possibly leading to incorrect conclusions.

4.2 Gathering method of onion addresses

As mentioned in Subsection 2.1, there are two main ways to collect onion addresses. Either by using a web crawler or by using the DHT approach.

A web crawler can only find onion addresses that are on any web page. These are often links to other websites. Onion addresses of services like SSH, FTP, email, chat or botnet infrastructures are not usually given on web pages and can therefore not be detected by web crawling. Avarikioti et al. describe the part of the Tor darknet that can be reached by web crawlers (the dark web) as “visible” [5]. These are onion services that respond to an HTTP(S) request on Port 80 or 443. This part should make up at least 15% and at most 50% of the entire Tor darknet. All other onion addresses are either offline, were reachable on other ports and provide other services or belong to botnets (see Subsection 4.4).

For a web crawler, a good selection of starting points (seed pages) is important. A theoretical way would be a list of randomly generated onion addresses. Since a v2 address is 16 characters long, the probability to find of a randomly generated active onion address is 2^{-65} if there are, for example, 30,000 active onion services [7]. Bernaschi et al. initially followed this approach, but found that of the approximately one million randomly generated addresses (calculated within a month), not a single one was active [7]. Therefore, this approach is not useful. Since the new v3 addresses are 56 characters long, the probability of randomly generating an active onion address decreases enormously. For this reason, many works use as seed

(1) lists compiled by third parties, such as onion.link [2], onion.cite [1, 17], FreshOnion [26, 27], hiddenWikis [6, 7, 11, 13, 19] or other clearnet sites (e.g. pastebins or reddit) [19, 23],
(2) existing darknet search engines (e.g. Ahmia or Ichidan) [1, 2, 13, 17, 26, 27], or
(3) onion lists from previous works [2, 5].

Especially in the case of (1), this can lead to a large bias, as these lists include a preselection of the party who compiled the list. Owensen et al. showed that a web crawler using such public seed lists finds less than half of the websites that can be found using the DHT

Work	Year [Published]	Method	Addresses found	Online	Analyzed	Content categorized	Permanent services	Size
Biryukov et al. [8]	02/2013 [2014]	DHT	39,824	24,511	21,325*	1,813	-	-
Spitters et al. [24]	07-12/2013 [2014]	crawler	>7,000	5,725	1,481	1,481	-	-
Owen and Savage [18]	2014 [2016]	DHT	80,000	~50,000*	n/a	n/a	~15% (>6m)	45,000
Moore and Rid [17]	01-03/2015 [2016]	crawler	5,615	5,205	2,723	2,723	-	-
Intelliagg [14]	02/2016 [2016]	crawler	29,532	13,585*	13,585*	13,585*	-	30,000 [DW]
Gollnick and Wilson [12]	08/2016 [2016]	crawler	n/a	n/a	n/a	400	-	-
Sanchez-Rola et al. [23]	05-06/2016 [2017]	crawler	198,050	7,257	7,257	7,257	-	-
Al-Nabki et al. [1]	05-07/2016 [2017]	crawler	>250,000	7,931	6,831	6,831	-	-
Avarikioti et al. [5]	n/a [2018]	crawler	34,714	10,957	7,566	7,566	-	-
Owenson et al. [19]	n/a [2018]	DHT	n/a	n/a	n/a	-	~40% (>2w)*	<50%
Yoon et al. [27]	01-07/2017* [2019]	crawler	100,000	13,326	13,326	-	-	-
Al-Nabki et al. [2]	05-07/2017 [2019]	crawler	124,589	3,536	3,536	3,536	-	-
Steinebach et al. [25]	2018 [2019]	DHT	173,190	82,145	60,036	-	~50% (>2m)	<50%*
Faizan and Khan [11]	n/a [2019]	crawler	25,742	6,227	4,102	4,102	-	-
Aoki and Goto [4]	06/2018-01/2021 [2021]	crawler	n/a	n/a	172,740	n/a	~42%** [DW]	40,848 [DW]

* Value has not been explicitly reported in the paper, but could be read or calculated by given statements, tables or graphs.

** Average value of all matched onion addresses of two 6-month time periods.

[DW] Value refers to the dark web, i.e. web services only.

Table 1: Comparison of previous researches

approach. However, the websites found by the web crawler account for 98% of visits [19].

When collecting onion addresses via DHT, onion services other than web services are also found. In addition, a seed list is not necessary and thus there is no bias. However, collecting via DHT requires time (or enormous resources). During the collection, all onion services that publish their descriptors at the observation period are found, regardless of the lifetime of the onion service itself. Thus, a lot of very short-lived onion services are collected (see Subsection 4.3). If these are all crawled and evaluated, then these short-lived onion services are overrepresented, because they were not online at the same time, but spread over a certain period of time [18].

4.3 Short lifetime of services

Various researchers mentioned that the Tor network has a variety of short-lived services.

When collecting onion addresses by crawling the dark web, the number of onion addresses found is much higher than the number of onion addresses that can be reached for analysis afterwards. On web pages on the dark web, addresses can be found that are already offline. For example, a study noted that more than half of all collected onion addresses - found by crawling the dark web - were not accessible during an analysis period of two weeks [14]. Al-Nabki et al. found over 250,000 onion addresses but only 7,931 were alive for analyzing [1]. In another work Al-Nabki et al. collected further 124,589 new onion addresses of which 3,536 were alive [2]. According to Avarikioti et al., within one year, up to 92.5% of onion addresses can go offline. They describe that for their work they used a list of 20,000 onion addresses from a previous year, of which only 1,500 were online [5]. This 1,500 onion addresses were used as seed for their own crawler to find further 34,714 addresses, but only 10,957 of them were reachable afterwards. A web crawler was also used by Yoon et al. [27]. They found 100,000 onion addresses in a time period of seven month but could only analyze 13,326 online onion addresses.

When collecting addresses via DHT it is possible to collect currently available onion addresses. These are onion services which descriptors are published during the observation. Biryukov et al. could find 39,824 onion addresses by collecting descriptors [9]. However, only 24,511 addresses were available afterwards for analysis [8]. Owen and Savage found in their work, that many onion services existed for a short period of time. Of 80,000 onion services, only 12,000 (15%) were alive after 6 months. After 18 months, fewer than 10,000 were still online [18]. Steinebach et al. estimate that only half of all services are permanent services with a lifetime longer than two month. Another work showed that 30% of all services were never accessible, although the respective descriptor had been published [19]. After 24 hours, over 50% of the observed services were no longer accessible.

As mentioned in Subsection 4.2, short-lived services can become a problem in the evaluation, as they may be overrepresented. The values of Tor Metrics as shown in Figure 1 are also overestimated [19]. The cumulative total number of onion services are recorded in any 24-hour period, and not on a point sample. Owenson et al. pointed out that the number of onion services is overestimated by a factor of two or more [19]. Accordingly, there should be fewer services than the numbers presented at Tor Metrics.

4.4 Botnet command and control servers

Due to the fact that Tor helps to make servers harder to locate, it has become popular for command and control (C&C) infrastructures of botnets [18]. As mentioned in Subsection 4.2, such services cannot be discovered by crawling the dark web. Most of all services, that a crawler cannot reach belong to botnet command and control addresses [19]. By using DHT, all onion addresses can be found from all types of services, such as SSH, FTP, email, chat or even botnet infrastructures. For a categorization of addresses for the previously mentioned services, usually the open ports of an onion service are analyzed [8, 19, 25]. Botnets usually use port numbers that do not correspond to the standardized port numbers of other services. Therefore, it is possible, but not necessarily reliable, to identify botnets by their port numbers.

In a 2013 study by Biryukov et al., over 50% of all analyzed onion addresses belonged to a botnet called "Skynet" [8]. The investigation of the most requested onion addresses confirmed this. Most of the requests led to addresses that belonged to either the botnet called "Skynet" or the botnet "Goldnet". In 2018 over 27,000 of 53,466 permanent services (lifetime over 2 month) belonged most likely to a botnet [25]. Botnets can consequently represent a large proportion of all onion services. However, categorization based on port numbers solely is not necessarily reliable. The botnet Goldnet was accessible under port 80 even if a 503 server error was returned [8]. In order to avoid incorrectly assigning an onion address to a service (e.g. a web service), every address would have to be checked for a possible botnet. As shown in the following subsection, many web pages have undetermined content. It is possible that botnets could potentially contribute to such a behavior.

4.5 Web services with undetermined content

In addition to classifying services, analysis regarding the content and its distribution represent a popular approach to producing descriptives concerning this context. To this end, previous studies examined and categorized the content of crawled websites [1, 2, 5, 8, 11, 12, 14, 17, 18, 23, 24]. Most of them found that only a subset of accessible websites could be evaluated. Many onion services offered no or insufficient website content to categorize. Many websites

(1) have content under 20 words [1, 2, 8, 11, 24],
(2) contain an error message embedded in an HTML page [5, 8, 11, 24],
(3) provide just images without any text [1, 2, 11],
(4) provide just an empty HTML page [2, 5, 17, 23]
(5) display a default web page of any service [8, 17, 23, 24],
(6) contain redirection links [11, 23, 24], or
(7) have other unreadable text or words [1].

Researchers excluded these websites from their categorization [5, 8, 11, 24] or placed them in categories such as "empty" or "none" [1, 2, 17]. Various studies report that English is the most common language on the dark web [2, 5, 8, 11, 14, 23, 24]. Of the studies that performed a content analysis, many exclude websites containing language other than English [1, 2, 5, 8, 26]. For reasons such as listed above, Biryukov et al., for example, were only able to categorize 1,813 out of 24,511 web services [8].

4.6 Duplicates of onion services

Another big issue in determining the distribution of web content are duplicates of websites. Researches show that the number of duplicates of websites in the Tor network are quite high. In 2017, Al-Nabki et al. showed, that the majority of "illegal suspicious" onion services are duplicates and are reachable under different onion addresses [2]. For example, 40% of all websites which were labeled as "drugs" and 90% of the category "cryptolocker" are duplicates. Yoon et al. [27] also showed that services such as bitcoin mixers and darknet marketplaces have a very high number of phishing websites in 2017. For example, the former mixer service *Bitcoin Fog* had 276 duplicates as fake websites. 165 phishing sites were found for the darknet marketplace *AlphaBay*. Barr-Smith and Wright investigated the amount of phishing onion web services within the Tor network in 2020 [6]. They analyzed 11,533 onion addresses of which over 51% were imitations of other websites. Brenner et al. showed that 20% of 258 single vendor shops are duplicates [10]. 31% have such high similarities to each other that they could be divided into seven similarity groups. The authors suspect that all shops in a similarity group belong together in some way [20]. The same vendor could be active in several shops or the websites were faked and misused as phishing sites.

A special case of duplicates are mirrors. Mirrors differ from phishing sites due to the fact that they are duplicated by the owners of a web service. These mirrors are supposed to make the offered services more resistant to DDoS attacks [6]. Furthermore, mirrors can also be used against bottlenecks in the network and to improve load balancing [27]. Yoon et al. showed the number of authentic onion addresses and the number of phishing addresses using two darknet marketplaces and three bitcoin mixer sites as examples [27]. Both marketplaces had multiple authentic onion addresses. This is an important information when analyzing the actual supply on the Tor network.

5 CASE STUDY - DARKNET MARKETPLACES

In order to demonstrate the difficulty of obtaining an accurate statement to describe the *darknet*, we use darknet marketplaces in the Tor network as a case study. In Subsection 5.1 we analyze the distribution of mirror sites of the current darknet marketplaces. Subsection 5.2 illustrates the difference in size in regards to the number of onion addresses of darknet marketplaces that can be found via web crawling versus the number of unique marketplaces that actually exist.

5.1 Mirrors of darknet marketplaces

This subsection analyzes the distribution of mirror sites using the example of all current darknet marketplaces. We identify platforms where darknet marketplaces and its main onion addresses are listed. Some platforms also present mirror addresses. This approach has been used by previous work to find authentic addresses of services [6, 27]. Two of common platforms for onion addresses and mirrors are *Darknet Live* or *dark.fail*. Both can be found on the clearnet, but also on the Tor network (darknet). Table 2 shows all platforms that present current darknet marketplaces where we collected their onion addresses in January 2022. At all we could find 37 darknet marketplaces and 184 onion addresses.

In order to check all found onion addresses for authenticity, we called up each marketplace manually and compared the collected addresses with the mirror addresses presented on the marketplace website. Of the 157 active onion addresses found, 145 were named as authentic main or mirror addresses on the respective marketplaces. Three addresses were redirected directly to an authentic confirmed address. Two addresses were not specified as authentic, however, we received the identical HTTP response headers as those for the authentic addresses. In addition, both addresses have the same backend as the authentic addresses. We were able to log in to both addresses with an username that we had registered on an authentic address. Both have been used in previous work to identify phishing sites [6, 27]. Seven addresses could not be verified for authenticity. These addresses belonged to the marketplaces Brightstar Fountain and Phoenix Market. One address belonged to Tor2door Market.

Table 3 shows all marketplaces and the number of found onion addresses through which these marketplaces can be reached. *Online* shows the number of active addresses during the analyzing process. Some marketplaces provide separated onion addresses that lists all official and authentic mirror addresses [27]. The column *ML* shows the number of active onion addresses that present mirror addresses for the respective marketplace. *Total* sums all active onion addresses that belong to a marketplace.

Platform	Type	DN / CN
dark.fail	Status Page, List	DN, CN
DarkEye	Status Page, List	DN
dark dot net	Darknet News, List of DM	CN
DarknetLive	Darknet News, List	DN, CN
Darknetstats	Darknet News, List of DM	CN
DarkNet Trust	DM Search Engine, List of DM	DN
DeepOnionWeb	Status Page, List	CN
Dread	Darknet Forum	DN
OnionLive	Status Page, List	CN
Raptor.life	Darknet News, List of DM	DN
recon	DM Search Engine, List of DM	DN

DN: Darknet. CN: Clearnet. DM: Darknet Marketplace
List: List of onion services

Table 2: Platforms with listed darknet marketplaces

As shown in Table 3 only 12 marketplaces have just one single onion address. All other marketplaces are accessible by multiple addresses. The Majestic Garden is reachable under ten different onion addresses. Further ten addresses belong to this marketplace as list of mirror addresses. According to this table, all darknet marketplaces have an average of four or five onion addresses for their website and mirror lists.

5.2 Insight into the total number of darknet marketplaces

It is difficult to determine the exact proportion of websites belonging to marketplaces. Table 3 shows all active multi-vendor marketplaces in January 2022. On these marketplaces multiple vendors can sell their products and services on different categories over these platforms and get rated by their users. Other selling platforms are single vendor shops [20]. These web shops have only a small number of web pages and only one vendor offers its products on this sites. There is no option available to rate the shop, the products or the vendor. This makes it difficult for Tor users to classify this shops as trustworthy.

Past works have shown that categorizing websites is not trivial. Some works categorize all websites with drug-related content into the category "drugs" [2, 5, 11, 18]. It doesn't have to be primarily darknet marketplaces or vendor shops. Marketplaces that do not deal with drugs are sorted into categories such as "marketplaces" or "markets". Others studies categorized website topics such as "drugs", various "counterfeit" categories (e.g. credit cards, money, personal identification) or "services". They do not have categories such as "marketplaces" [8, 12, 14, 17, 26]. Still others do not subdivide such websites at all and sort all these websites under the category "market/shopping" [23]. While various works have shown that marketplaces are not the largest category in their respective data sets [5, 14, 17, 23], other works show that the largest categories are darknet marketplaces or drugs [2, 8, 18]. Faizan and Khan describes that drug-related websites make up between 5% and 15% of all websites on the dark web across various works [11]. They explain this differences by the high turnover of services and websites within the darknet ecosystem.

Name	Onion addresses	Online	ML	Total
The Majestic Garden	10	10	10	20
Alien Market	18	18	0	18
Tor2door Market	9	8	2	10
ARES Market	10	10	0	10
World Market	8	8	0	8
Revolution Market	14	6	0	6
CannaHome	4	4	2	6
DarkFox Market	6	6	0	6
Quest Market	6	6	0	6
MGM Grand Market	1	1	4	5
Royal Market	5	5	0	5
ColombiaConnection Market	5	5	0	5
ASAP (old: ASEAN)	9	4	0	4
Brightstar Fountain	4	4	0	4
Incognito Marketplace	4	4	0	4
Kingdom Market	4	4	0	4
BlackHole Market	3	3	0	3
HeinekenExpress	3	3	0	3
Retro Market	3	3	0	3
Vice City Market	3	3	0	3
WeAreAMSTERDAM	3	3	0	3
Versus	7	2	0	2
Bohemia Market	2	2	0	2
Dark0de Reborn	2	2	0	2
Phoenix Market	2	2	0	2
UnderMarket 2.0	1	1	1	2
Abacus	1	1	0	1
AlphaBay Market	1	1	0	1
Babylon Market	1	1	0	1
Cocorico Market	1	1	0	1
Cypher Market	1	1	0	1
Digital Thrift Shop	1	1	0	1
Hermes Market	1	1	0	1
Hydra	1	1	0	1
Nemesis Market	1	1	0	1
Silk Road 4	1	1	0	1
WeTheNorth Market (WTN)	1	1	0	1
Total	157	138	19	157
∅	4.24	3.73	0.51	4.24

Table 3: Current darknet marketplaces and number of onion addresses

In order to find out how many websites can be found that belong to a darknet marketplace compared to the actual number of darknet marketplaces, we crawled 10,000 onion websites in 2020. In a randomly compiled data set of 5,000 websites, we identified 558 addresses belonging to darknet marketplaces. We extracted and analyzed the titles of the websites in order to assign them to darknet marketplaces. We examined all 558 onion addresses for identical websites and removed all duplicates. After that, only 293 onion addresses remained as seen in Figure 2. With the remaining 293 addresses, we determined that 244 addresses are merely vendor shops. Due to the lack of reputation options on these vendor shops, it is hardly to check the authenticity of these shops for users. Of the remaining 49 onion addresses, only 31 were online at the time of the analysis in December 2020.

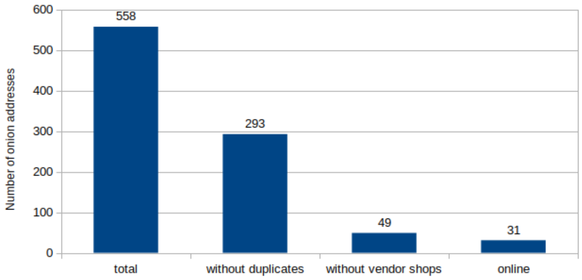

Figure 2: Number of onion addresses belonging to darknet marketplaces

We did not verify the authenticity of the remaining 31 onion addresses. However, this evaluation shows that a very high number of darknet marketplaces are identified by crawling websites. As noted by other works, there are many phishing websites [6, 27]. Subsection 5.1 shows that many marketplaces provide duplicates of their service as mirrors. All onion addresses of both types of duplicates are found through web crawling. For this reason, many works get quite a high number of websites in a certain category. In comparison, the actual offer for a Tor user is much lower.

6 CONCLUSION

Interest in gaining information concerning *the darknet* and in particular Tor stems from various parts of society. However, depictions from various societal contexts appear to be inconsistent and at times implausible. To this end, we reviewed a sample of studies for critical aspects concerning research in the context of Tor. In particular, we examined these studies with respect to (1) the terminology, (2) the methodology to collect the sample and (3) the analysis of the data. Our results indicate that critical aspects in particular pertain to the following: (A) An inconsistent use of terminology, (B) the methodology with the sample was gathered, as well as the handling of (C) short-lived services, (D) botnet command and control servers, (E) web services with undetermined content and (F) duplicated of onion services. In the following, we will conclude by summarizing the implications of the handling of these aspects as we identified them.

An inconsistent use of terminology results in problems concerning the robustness of the research. This is for example the case if an inadequate terminology is used to describe the context of the research (e.g. darknet, dark web, deep web, etc.) to which the results refer to. Additionally, comparability of results for further work is also jeopardized.
The analyzed works show two prominent methods for collecting a sample of onion addresses. The way in which a sample is compiled is crucial to the research, as it serves as the basis for any subsequent analysis. Therefore, it is important that any analysis based on the sample is related to how the sample is composed. Pertaining both methodologies to compile a sample in the context of research on Tor, they implicate substantial differences. Depending on the method applied to find onion addresses affects the scope of onion addresses that may be found. Hence, results concerning, for example, the size of Tor or the amount of onion addresses found, must be reflected in respect to the methodology used to compile the sample. The consequences of this relate in particular the aspect of short-lived services. The number of offline services must be reflected with respect to the method of how the sample was collected. When a sample is collected via DHT, we can presume that all services were available at the point of data collection. Thus, the time between collecting the addresses and analyzing how many of those are actually online, is vital. To gain a better understanding of the fluctuation and share of short-lived services a longitudinal design (i.e. repeated measurements) seems promising.

Another challenge concerning statements with relation to *the size* of Tor involves botnet command and control servers. If the size of Tor is measured through the amount of onion addresses, this might include a large share of botnets. Therefore, it is not wrong to include these numbers when reporting the amount of onion services. However, several thousand onion addresses may belong to a single botnet. Therefore, the total number of available onion addresses does not necessarily represent the number of unique onion services in the Tor network. It is crucial to reflect these results in terms of their implications. When examining onion services in terms of content, it is crucial to check the exclusion criteria of the analysis. This way, a better understanding can be built concerning the basis of the analyzed data. While numerous studies indicate portions of onion services as web services of undetermined content, this can be attributed to a variety of reasons, such as how the exclusion criteria are decided. For example, some studies apply limitation to language, type of content (e.g. images versus text, minimum number of words, etc.). Exclusion criteria may be conditioned, for example, by the method applied to analyze the sample.

Finally, a common narrative consists of categorizing a large portion of onion services as marketplaces. However, as we also demonstrated in our case study, duplicates of onion services must be considered in the initial categorization. When this is not considered, it leads to distorted reporting of the distribution of web service content. Moreover, it may imply a crime and threat situation that does not accurately reflect the reality of Tor.

ACKNOWLEDGMENTS

The joint project PANDA on which this publication is based was funded by the Federal Ministry of Education and Research under the funding codes 13N14355 and 13N14356. The authors are responsible for the content of this publication.

REFERENCES

[1] Mhd Wesam Al Nabki, Eduardo Fidalgo, Enrique Alegre, and Ivan de Paz. 2017. 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*. 35–43.

[2] Mhd Wesam Al-Nabki, Eduardo Fidalgo, Enrique Alegre, and Laura Fernández-Robles. 2019. Torank: Identifying the most influential suspicious domains in the tor network. *Expert Systems with Applications* 123 (2019), 212–226.

[3] Abdullah Alharbi, Mohd Faizan, Wael Alosaimi, Hashem Alyami, Alka Agrawal, Rajeev Kumar, and Raees Ahmad Khan. 2021. Exploring the topological properties of the Tor Dark Web. *IEEE Access* 9 (2021), 21746–21758.

[4] Taichi Aoki and Atsuhiro Goto. 2021. Graph visualization of dark web hyperlinks and their feature analysis. *International Journal of Networking and Computing* 11, 2 (2021), 354–382.

[5] Georgia Avarikioti, Roman Brunner, Aggelos Kiayias, Roger Wattenhofer, and Dionysis Zindros. 2018. Structure and content of the visible Darknet. *arXiv preprint arXiv:1811.01348* (2018).

[6] Frederick Barr-Smith and Joss Wright. 2020. Phishing With A Darknet: Imitation of Onion Services. In *2020 APWG Symposium on Electronic Crime Research (eCrime)*. 1–13. https://doi.org/10.1109/eCrime51433.2020.9493262

[7] Massimo Bernaschi, Alessandro Celestini, Stefano Guarino, and Flavio Lombardi. 2017. Exploring and analyzing the tor hidden services graph. *ACM Transactions on the Web (TWEB)* 11, 4 (2017), 1–26.

[8] Alex Biryukov, Ivan Pustogarov, Fabrice Thill, and Ralf-Philipp Weinmann. 2014. Content and popularity analysis of Tor hidden services. In *2014 IEEE 34th International Conference on Distributed Computing Systems Workshops (ICDCSW)*. IEEE, 188–193.

[9] Alex Biryukov, Ivan Pustogarov, and Ralf-Philipp Weinmann. 2013. Trawling for tor hidden services: Detection, measurement, deanonymization. In *2013 IEEE Symposium on Security and Privacy*. IEEE, 80–94.

[10] Fabian Brenner, Florian Platzer, and Martin Steinebach. 2021. Discovery of Single-Vendor Marketplace Operators in the Tor-Network. In *The 16th International Conference on Availability, Reliability and Security*. 1–10.

[11] Mohd Faizan and Raees Ahmad Khan. 2019. Exploring and analyzing the dark Web: A new alchemy. *First Monday* (2019).

[12] Clare Gollnick and Emily Wilson. 2016. Separating fact from fiction: The truth about the dark web. *Terbium Labs* (2016).

[13] Clement Guitton. 2013. A review of the available content on Tor hidden services: The case against further development. *Computers in Human Behavior* 29, 6 (2013), 2805–2815.

[14] Intelliagg. 2016. Deeplight: Shining a Light on the Dark Web. , 12 pages.

[15] George Kadianakis and Karsten Loesing. 2015. Extrapolating network totals from hidden-service statistics. *Tor Tech Report* 01 (001 (2015).

[16] Mihnea Mirea, Victoria Wang, and Jeyong Jung. 2019. The not so dark side of the darknet: A qualitative study. *Security Journal* 32, 2 (2019), 102–118.

[17] Daniel Moore and Thomas Rid. 2016. Cryptopolitik and the Darknet. *Survival* 58, 1 (2016), 7–38.

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

[19] Gareth Owenson, Sarah Cortes, and Andrew Lewman. 2018. The darknet's smaller than we thought: The life cycle of Tor Hidden Services. *Digital Investigation* 27 (2018), 17–22.

[20] Florian Platzer, Fabian Brenner, and Martin Steinebach. 2022. Similarity Analysis of Single-Vendor Marketplaces in the Tor-Network. *Journal of Cyber Security and Mobility* (2022), 205–238.

[21] Tor Project. 2022. Tor Metrics Unique .onion addresses. https://metrics.torproject.org/hidserv-dir-onions-seen.png?start=2013-01-01&end=2021-08-28. Accessed: 2022-02-07.

[22] Tor Project. 2022. Torspec dir-spec.txt. https://github.com/torproject/torspec/blob/main/dir-spec.txt. Accessed: 2022-05-10.

[23] Iskander Sanchez-Rola, Davide Balzarotti, and Igor Santos. 2017. The onions have eyes: a comprehensive structure and privacy analysis of tor hidden services. In *Proceedings of the 26th international conference on world wide web*. 1251–1260.

[24] Martijn Spitters, Stefan Verbruggen, and Mark Van Staalduinen. 2014. Towards a comprehensive insight into the thematic organization of the tor hidden services. In *2014 IEEE Joint Intelligence and Security Informatics Conference*. IEEE, 220–223.

[25] Martin Steinebach, Marcel Schäfer, Alexander Karakuz, Katharina Brandl, and York Yannikos. 2019. Detection and analysis of Tor onion services. In *Proceedings of the 14th International Conference on Availability, Reliability and Security*. 1–10.

[26] Sugiu Takaaki and Inomata Atsuo. 2019. Dark web content analysis and visualization. In *Proceedings of the ACM International Workshop on Security and Privacy Analytics*. 53–59.

[27] Changhoon Yoon, Kwanwoo Kim, Yongdae Kim, Seungwon Shin, and Sooel Son. 2019. Doppelgängers on the dark web: A large-scale assessment on phishing hidden web services. In *The World Wide Web Conference*. 2225–2235.