Unveiling the I2P web structure: A connectivity analysis

Roberto Magán-Carrión, Alberto Abellán-Galera, Gabriel Maciá-Fernández and Pedro García-Teodoro
Network Engineering & Security Group
Dpt. of Signal Theory, Telematics and Communications - CITIC
University of Granada - Spain
Email: rmagan@ugr.es, albertoabellan@correo.ugr.es, gmacia@ugr.es, pgteodor@ugr.es

***Abstract*—Web is a primary and essential service to share information among users and organizations at present all over the world. Despite the current significance of such a kind of traffic on the Internet, the so-called Surface Web traffic has been estimated in just about 5% of the total. The rest of the volume of this type of traffic corresponds to the portion of the Web known as *Deep Web*. These contents are not accessible by usual search engines because they are authentication protected contents or pages only reachable through technologies denoted as *darknets*. To browse through darknet websites, special authorization or specific software and configurations are needed. TOR is one of the most used darknet nowadays, but there are several other alternatives such as I2P or Freenet, which offer different features for end users. In this work, we perform a connectivity analysis of the websites in the I2P network (named *eepsites*) aimed to discover if different patterns and relationships from those used in legacy webs are followed in I2P, as well as to get insights about the dimension and structure of this darknet. For that, a novel tool is specifically developed by the authors and deployed on a distributed scenario. Main results conclude the decentralized nature of the I2P network, where there is a structural part of interconnected eepsites while other several nodes are isolated probably due to their intermittent presence in the network.**

***Index Terms*—Deep Web, Darknet, I2P, Crawling, Eepsite, Connectivity.**

I. Introduction

The quintessential Internet service, the World Wide Web (WWW), simply known as “the Web”, has a hidden face which is ignored by a great number of users and organizations. The majority of them browses the so-called *Surface Web*, which corresponds to web resources able to be indexed by common search engines. However, the Surface Web is just the peak of an enormous WWW iceberg. The great part of the WWW contents are hosted on the so-called *Deep Web* [1]. These contents are not accessible by usual search engines mainly, because they involve authentication protected contents or pages that are only reachable through the well known *darknet* related technologies [2], [3]. To browse through darknet websites , a special authorization or specific software and configurations are needed, indeed.

Darknets have been widely publicized to users as technologies mainly intended to elude censorship restrictions imposed by totalitarian governments [4], and to preserve fundamental security rights like privacy or anonymity. Although this is true, the generalized use of darknets surpasses nowadays freedom demands to become a main tool for illegal actions and cybercriminality because of the anonymity provided by them [5].

Mainly motivated by that, darknets have received the attention of researchers in the last years. This way, some works in the literature have analyzed the content and services offered through this kind of technologies [6], [7], [2], as well as other relevant aspects like site popularity [8], topology and dimensions [9], or classifying network traffic and darknet applications [10], [11], [12], [13], [14].

Two of the most popular darknets at present are *The Onion Router* (TOR; https://www.torproject.org/) and *The Invisible Internet Project* (I2P; https://geti2p.net/en/). This paper is focused on exploring and investigating the contents and structure of the websites in I2P, the so-called *eepsites*. In particular, we are interested on how they are interconnected and which kind of relationships exist among them. Although some works have been focused on TOR [15], [16] in this research line, no efforts exist regarding I2P. In fact, we only found the work [17], where the authors attempt to discover and analyze eepsites, they claiming to uncover the 80% of the total eepsites in I2P. However, no eepsites' relationships and connectivity analysis are provided in this work.

To analyze the structure and connectivity of websites in I2P, a crawling and scrapping tool has been specifically developed here: `c4i2p` (crawling for I2P), which is able to extract some useful information from eepsites to characterize them and to learn how they are inter-connected. For that, a three months long experiment has been carried out in a distributed environment composed by different nodes located at different places. As we will discuss below in the document, around only $\sim 1.5\%$ of the total observed services correspond to eepsites. Besides, they are generally small and simple, composed of few pages with not so many text and hypertext. Moreover, about $\sim 66\%$ of the eepsites are isolated from the rest, thus comprising a hidden part of the overall set of eepsites. That means that the I2P darknet can be seen as an heterogeneous and decentralized network, where also some popular sites exist. Such popular eepsites are characterized by their persistence in the I2P network over time, they constituting a kind of backbone for the eepsite network. The rest of them seem to be randomly appearing and disappearing in/from the network as it will be shownthrough the present work.

The rest of the document is organized as follows. A main background on analysis of darknet related technologies is provided in Section II. After that, Section III introduces the main fundamentals of I2P and presents the c4i2p tool aimed at discovering the structure of eepsites. The experimental setup using the proposed tool is detailed in Section IV, while the results extracted from the experimentation are afterwards discussed in Section V. Finally, the main conclusions and future work are presented in Section VI.

II. Background on darknets

The Surface Web or simply the Web, has been widely studied by the research community with different and heterogeneous aims. A number of works are focused on different topics, such as performance optimization for search engines [18], [19], [20], [21], analysis of connectivity, dimension and structure of the sites [22], [23], [24], [25] and classification of the sites and their contents [26], [27], [28]. As we will see in the following, works on the Deep Web and darknets have also addressed similar topics than in the Surface Web.

A. Hidden Services and Contents

An analysis of the popularity of hidden services in TOR is carried out in [6] and [9]. Both works are aimed at analyzing how much traffic is due to the use of hidden services and how many unique `*.onion` addresses exist in the network, in order to estimate the number of TOR hidden services.

In [29], a generic crawling tool is proposed to discover resources with different contents hosted both on the Surface Web or the Deep Web. In particular, the authors carry out an experiment intended to search websites with content about homemade explosives. Similarly, authors in [8] introduce an approach to enumerate all hidden services in TOR. For that, they take advantage of defects in both the design and implementation of TOR hidden services that could allow an attacker to deanonymize them.

In 2018, the creation of a database called DUTA was presented in [7]. The database contains a list of a multitude of TOR hidden services classified by content. It is noteworthy to notice that more than 250,000 addresses of hidden services were found, but only 7,000 of them were accessible while the rest were down or unavailable.

More recently and in the same line of the previous work, a new algorithm called TORank is proposed in [2] to classify hidden services in TOR. Authors thoroughly analyze site contents, then creating a dataset (DUTA-10K) which updates DUTA. In addition, ToRank is assessed and quantitatively compared with some of the most popular classification algorithms, such as PageRank, HITS and Katz. Among the most interesting conclusions we can mention is the fact that only 20% of the available hidden services refer to suspicious activities, while 48% of them are associated with normal activities. They also discovered that, in general, domains related to suspicious activities have multiple clones under different locations (URLs), which is susceptible to be used as an additional feature to identify them.

A week-long experiment was also carried out in [30] to monitor the I2P network. Some interesting results were obtained in this study. Among others, the authors conclude that 37% of the published leaseSets were offline after publication on NetDB. Around 30% of the complete leaseSets set were not identified, which means that 30% of the network was not running neither a web server nor an I2PSnark client. In addition, more than 50% of the anonymous web services discovered were 70% of the time available online. It should be noticed that the experiment had a relatively short duration and it was performed when the network was possibly not yet mature.

On the other hand, a system for crawling the Deep Torrent, that is the torrents available in BitTorrent that cannot be found through public sites or regular search engines, is presented in [31]. Authors estimate what percentage of resources shared in the BitTorrent network are hidden or part of the Deep Web.

B. Network Topology and Connectivity Analysis

Regarding network topological aspects, Liu *et. al* introduce in [32] two passive and active methods to discover I2P routers. An experiment was carried out for two weeks, where around 25,640 routers were discovered every day. The authors claim that they almost cover 94.9% of the I2P network compared to the data announced on the official website.

In [33], the authors make use of graphs to analyze the most popular emerging products in TOR. For that, authors make use of text information extracted from the domains corresponding to markets to create a Product Correlation Graph (PCG). In a similar line, a link inter-connectivity study is carried out in [16] about the structure and privacy of TOR hidden services. The authors analyze more than 1.5 million URLs hosted in 7,257 TOR domains. For each page, links, resources and redirection graphics, as well as the language used, are analyzed. Very relevant conclusions are extracted, such as the fact that domains in TOR are highly interconnected and that there are many links within TOR pointing to pages hosted on the Surface Web (the reverse case is also quite common).

In [15], a very complete survey about the topology and content of TOR is presented. Authors make use of a specific crawler that recursively explores the links found. Using it, a total of 34,000 hidden services are found, 10,000 of them corresponding to online services. The work concludes that most of hidden services are well connected through central sites such as wikis and forums. That is, the structure of this network is somewhat centralized. Regarding contents, half of the sites involve lawful activities, while those with illegal hidden services are mainly related with fraud, sale of counterfeit products and drug markets.

Another study intended to analyze I2P nodes is conducted in [34]. In it, authors collected 16,040 I2P nodes and analyzed some of their properties, such as distribution by country, bandwidth utilization and node FloodFill attributes. A similar work, and perhaps one of the most complete academic studies on I2P to date, can be found in [4]. This paper contributes a comprehensive empirical study of the network where population, cancellation rate (abandonment of the network by users), type of router and geographic distribution of the I2P pairs are measured. At the time of the study, there were about 32,000 active I2P pairs in the network per day, and 14,000 of them were behind NAT or firewalls. Moreover, despite the decentralized nature of I2P, a censor actor could block more than 95% of the peer IP addresses known by a stable I2P client by only controlling 10 routers in the network, what constitutes a serious deterioration of the connectivity of an I2P client.

C. Network Operation

An interesting study focused on measuring performance parameters for TOR can be found in [35]. There, response time, throughput and latency are measured, among others parameters. The authors conclude that browsing TOR is slower than doing it through the Surface Web.

Additionally, several works have been proposed based on the use of Machine Learning (ML) techniques and algorithms with the aim of analyzing and classifying darknet's network traffic in some sense. For instance, a ML-based approach to analyze traffic flows generated by I2P applications and users is introduced in [36]. The work concludes that it is possible to create both user and application profiles, and that the accuracy in creating such profiles depends on the amount of shared bandwidth. More recently, the authors in [10] introduce a hierarchical machine learning based framework to classify the type of network, the type of traffic and the applications that generated them. For that, a labeled dataset (Anon17) comprising network traffic collected from I2P, TOR and JonDonym darknets is used. In comparison to their previous work in [11], where they used a flat approach with several classic ML algorithm, they obtain now an improved performance specially in traffic classification. Moreover, the authors conclude that I2P applications are the hardest ones to classify in comparison to TOR.

With the same aim, the authors propose in [12] a three step XGBoost ML-based framework. As in the previous cases, the Anon17 dataset is used to validate the solution. A relevant improvement, in terms of classification accuracy, is achieved in comparison to other works like [36]. Moreover, they also tested the suitability of the approach for classifying different but similar network traffic coming from Virtual Private Networks (VPN) or intentionally encrypted. Similar works addressing the classification of I2P network data traffic can also be found in the literature [13], [14].

Some other works analyze the main weaknesses of TOR and I2P regarding anonymity that might compromise user identities and communication links. This is the case of [37], where the authors conclude that some of the attacks require considerable resources to be effective and that, therefore, it is very unlikely that they will succeed against this kind of networks. As a consequence, both darknets are considered highly safe.

A theoretical comparison between TOR and I2P is carried out in [38] from a number of perspectives: visibility by the community, scalability, memory usage, latency, bandwidth, documentation, vulnerability to DoS attacks, number of exit nodes, etc. Authors in [39] try to characterize the file-sharing environment within I2P, they evaluating how it affects the anonymity provided by the network. It is concluded that most of the activities within I2P are oriented to file sharing and anonymous web hosting. Moreover, it is also concluded that the nodes are geographically distributed.

The I2P network has also been analyzed from a forensic point of view in [40]. This work is aimed to help a forensic analyst to find clues related to the activity of cybercriminals in I2P. More deeply, a study on the performance and safety of I2P is presented in [41]. The authors compare the design of the NetDB with the design of the popular KAD (Kademlia) algorithm. Among other conclusions, they stand out that I2P users are, in general, more stable than those of other non-anonymous P2P networks. Because of that, KAD design is considered less vulnerable to different attacks than the current NetDB design is.

Through the above brief review of the state of the art, we can conclude that the Surface Web has been explored and analyzed to a large extent from many points of view and approaches. We have focused the discussion on research examples mainly based on crawling, information gathering, search engines and graph-shaped connectivity. Although to a lesser extent than the Surface Web, TOR has also been studied in depth, from several points of view too, e.g., content, structure (best result of this analysis is the TorMetrics Project, [42]), etc. Nevertheless, I2P is still somehow unknown and further works about its operation and structure are desirable, specially for eepsites. In [17], an attempt to discover and analyze eepsites is made. For that, authors propose the use of Floodfills, the active collection of `host.txt` files and the network crawling. A total of 1,861 online eepsites were discovered, this amount corresponding to 80% of eepsites existing in this network as the authors claim. However, no eepsites' relationships and connectivity analysis is provided. Because of that, the current work is mainly focused on the study of the I2P eepsites' relationships and interconnections to unveil the I2P web structure.

III. I2P Darknet: Basics and Analysis

In this section, we first present some fundamentals of I2P and afterwards describe the specific crawler developed to analyze that darknet.

A. *I2P fundamentals*

The I2P project [43] was started in 2003 by a group of hackers, developers and software architects as a useful set of tools against censorship, with anonymity and privacy in mind.

I2P operation is similar to that in other anonymous networks. Thus, traffic is routed through various points in the network using chains of proxy servers. Each data packet is anonymously routed to the corresponding destination. Moreover, when users register their I2P instances (routers), these become *nodes* of the network, thus contributing with their own network bandwidth to transport communications among other users.

The use of I2P is accomplished by using applications specifically designed to interact with I2P in a transparent way. Two relevant examples of that are I2PSnark (client for BitTorrent), and I2PTunnel (services to route TCP/IP flows for common applications such as web browsers, IRC or SSH), among others.

One of the main elements that conform the I2P architecture are tunnels, since they allow sending (outbound tunnel) and receiving (inbound tunnel) messages between users/nodes in the network. A tunnel could be defined as the composition of a set of routers that are in charge of sending and receiving data packets to and from specific destinations or sources.

I2P makes use of a simple combination of symmetric and asymmetric encryption algorithms to provide data confidentiality and message integrity. It is commonly named as *garlic* routing or encryption. More specifically, ElGamal/AES + SessionTags are used for the encryption with 2048-bit

ElGamal, AES256, SHA256, and 32-byte nonces (single-use numbers).

The Network Database or NetDB is an important part of I2P as well. NetDB is a Distributed Hash Table (DHT) with a structure based on the Kademlia algorithm. It contains the information about I2P services so that nodes in the network can communicate with each other. NetDB is queried the first time a running instance of I2P wants to contact another router in the network. NetDB is disseminated by means of the technique called *Floodfill*, which is based on the use of a series of special routers with the same name. Such special routers are in charge of distributing the database and synchronizing the corresponding information about the *routerInfo* and the *leaseSets* found on the network [44]. Unlike TOR directory authorities, the Floodfill routers that compose NetDB are not fixed or trusted routers. Indeed, any router on the network can become a Floodfill if it is configured in that way.

B. C4i2p: A Crawling Tool for the I2P Darknet

Although some crawling tools have been developed to date with different aims [45], none of them cover completely the main objectives this work is intended for: a connectivity analysis among eepsites. For this reason, a new and ready-to-use tool has been devised and developed: `c4i2p`, which stands for 'crawling for I2P'. Beyond its usefulness to achieve the objectives of the current work, we are convinced about its potential use by the research community to be improved or extended in many ways. Because of that, c4i2p has been conceived as an open source project and thus it is publicly available for downloading at [46].

Through the rest of this section, the main modules, processes and components that shape the developed tool are introduced and described.

1) Modules: As mentioned before, c4i2p tries to be a specific tool that allows users to interact with the I2P network. Thanks to this tool, relationships among eepsites can be observed by gathering different kind of information. For that, several modules are involved in the operation of c4i2p, as shown in Fig. 1:

- **Management**. It is in charge of controlling the entire workflow of the tool. It governs the interaction between the rest of procedures and elements composing the system. The management module is in turn managed by the end user.
- **Discovery**. It is responsible for checking the availability of the observed eepsites in a controlled manner over time.
- **Crawling**. It is responsible for carrying out the task of crawling and everything related with searching for and scrapping specific information from I2P eepsites.
- **Data Storage**. It is a module aimed to provide data storage services to others. It is responsible for the storage, persistence and management of all the information.

2) Processes, components and relationships: The main components that intervene in the operation of c4i2p are as follows (see Fig. 1):

- **Sources**. They are eepsite addresses from I2P. They can be obtained from three different origins:
  - SEED. It is obtained from the list of initial seeds (`seeds.txt`) from which the experiment starts. They were selected from a preliminary search for eepsites in I2P.
  - FLOODFILL. They are discovered by the Floodfill I2P deployed routers.
  - DISCOVERED. The ones discovered through the crawling process, i.e. from links obtained from each scrapped eepsite.
- **Manager**. The main procedure of the Management module, the Manager is the brain and core of the entire tool. It is in charge of coordinating and orchestrating the rest of elements and to keep track of eepsite's status and lifecycle.
- **Discoverers**. As the main part of the Discovery module, they implement the main functionality of the discovery procedure. Each one is responsible for checking if an eepsite is active (reachable and operational). How many of them and when they are launched is controlled by the Discovery module.
- **Spiders**. Each of them, separately, is in charge of carrying out the Crawling process of a certain *eepsite*. An spider dives recursively into all the pages to extract the desired information. It is launched right after an eepsite is discovered.
- **Database**. The database stores in an ordered and structured way all the information extracted from eepsites, as well as the inherent data of the system itself.
- **I2P network**. The I2P network is where the eepsites are hosted, which are the entities from which we want to extract information.
- **I2P router**. Finally, the I2P router allows c4i2p to communicate with the I2P network.

3) Eepsites functional states: During its lifecycle in c4i2p, an eepsite goes through several states. The states and transitions among them are depicted in Fig. 2:

- **DISCOVERING**. When c4i2p takes a new eepsite from the list of sources to check, the target eepsite is in process of being discovered. This is the initial state of every eepsite from the perspective of c4i2p, so that this state is concluded when the eepsite is successfully contacted or an error/timeout occurs.
- **PENDING**. Once an eepsite is successfully contacted, i.e., it is reachable and available, it goes to this state where it is awaiting to be crawled. Indeed, the eepsite is queued into a FIFO based queue for that.
- **ONGOING**. The eepsite is being crawled.
- **ERROR**. An eepsite is in the ERROR state if an error appears while crawling it. In order to avoid not contemplated errors during the crawling process, the application allows a certain number crawling processes until the eepsite is proposed to be discovered again.
- **FINISHED**. A node is labeled as FINISHED if it was successfully crawled.
- **DISCARDED**. An eepsite reaches this state after a certain number of discovering attempts or when a predefined on-discovering time is exceeded. This state concludes that the eepsite is no longer available or it never existed.

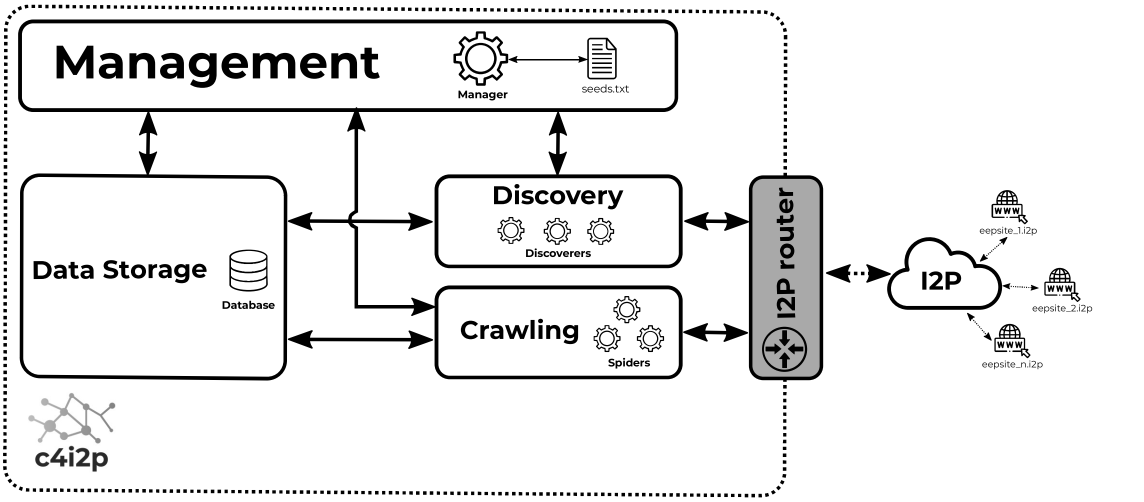


Figure 1. C4i2p modules and interactions.

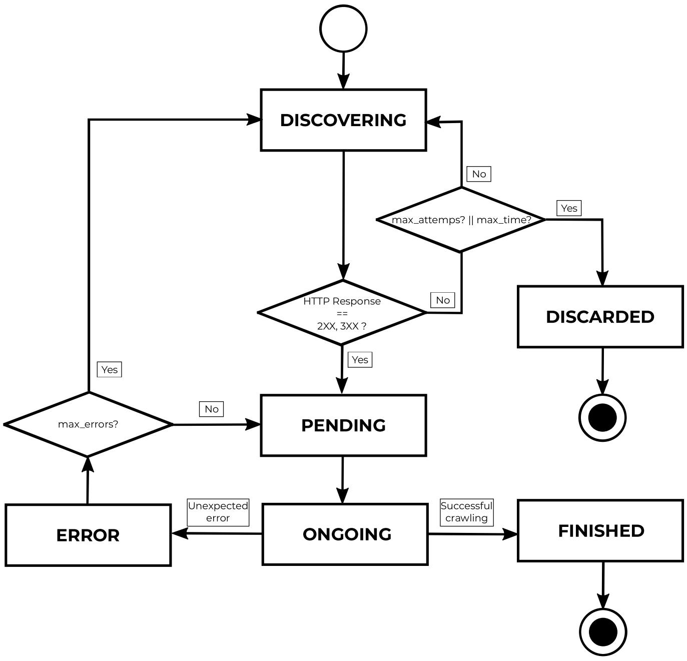


Figure 2. Eeepsite's lifecycle in c4i2p, where states (squared boxes) and available transitions (polygonal shapes) among them are shown.

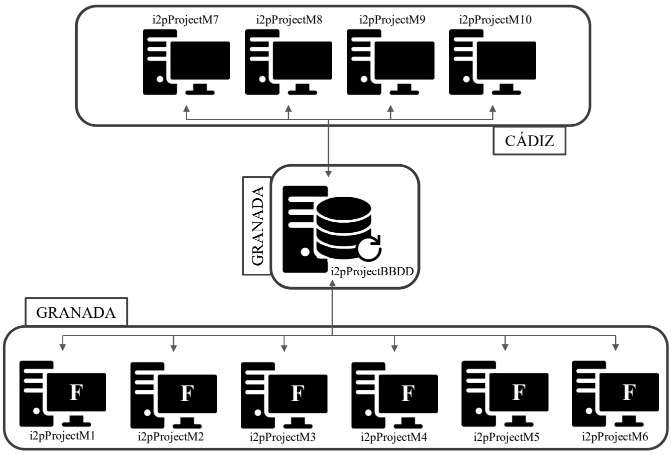


Figure 3. Experimental setup where 11 virtual machines are deployed and distributed in different locations: `i2pProjectM[1-6]` and `i2pProjectBBDD`, in Granada (Spain); and `i2pProjectM[7-10]`, in Cádiz (Spain). All the machines (except that corresponding to BBDD) run an I2P router. The ones with IDs `i2pProjectM[1-6]` are configured as floodfill I2P routers.

IV. Experimental setup

In order to crawl the I2P network a distributed experimental environment is devised. It is composed of 11 virtual machines organized in two separated groups, as shown in Fig. 3. They are distributed in different geographical places running in different computation clusters too. On the one hand, 7 of the VMs are deployed on the facilities of the University of Granada (Spain): 6 named as `i2pProjectM[1-6]` in the figure, and the one acting as the central database identified as `i2pProjectBBDD`. On the other hand, the remaining 4 VMs are deployed on the facilities of the University of Cádiz (Spain), which are identified in the figure as `i2pProjectM[7-10]`. All the virtual machines run Linux Ubuntu 18.04.1 LTS distribution with a total of 5 GBs RAM, 100 GBs HD and two 2.40 GHz vCPUs.

Every VM executes an I2P router instance for accessing the I2P darknet, except for the `i2pProjectBBDD` machine. Machines `i2pProjectM[1-6]` were set up as floodfill I2P routers, while the rest `i2pProjectM[7-10]` run as normal I2P routers. Floodfill routers, as mentioned in Section III-A, allow to discover additional eepsites since they all are in charge of maintaining and managing the NetDB. Following recommendations from work [47], we set up shared router's communication bandwidth to 8 MBps, the number of maximum participants in tunnels being equal to 10K. In addition, it is necessary to activate Floodfill routers by setting up `router.floodfillParticipant=true` in the `router.config` configuration file.

All VMs also include a c4i2p instance. They altogether provide a distributed crawling procedure where each c4i2p instance is in charge of managing and discovering its own eepsites. Both, Discovery and Crawling modules were accordingly set up in each c4i2p instance. Among other configuration parameters, the number of simultaneous running spiders is

Parameter	Value	Description
`MAX_ONGOING_SPIDERS`	10	Number of simultaneous spiders
`MAX_CRAWLING_ATTEMPTS_ON_ERROR`	2	Number of crawling attempts per eepsite
`MAX_DISCOVERY_ATTEMPTS`	30d × 24h	Number of discovery attempts per eepsite: one per hour during a month
`MAX_DISCOVERY_DURATION (m)`	30d × 24h × 60m	Maximum time for which an eepsite is tried to be discovered
`MAX_DISCOVERY_SINGLE_THREADS`	50	Number of eepsites to be simultaneously discovered
`HTTP_TIMEOUT (s)`	30s	HTTP request timeout from an I2P URL
`INITIAL_SEEDS`	`seeds.txt`	Path to the file containing the set of eepsite URLs
`INITIAL_SEEDS_BATCH_SIZE`	∼394	Number of initial seeds assigned to a c4i2p instance

Table I
C4I2P CONFIGURATION PARAMETERS.

limited to 10 to efficiently manage the VM computation capacities. Moreover, a maximum value is established for the discovery time and the number of attempts. If one of such limits is reached, the site in process of being discovered will be discarded. At the beginning, some initial I2P URLs (called *seeds*) are equally distributed among all c4i2p instances. In our experiment we initially count on 3,938 seeds, which means that ∼394 initials seeds are assigned to each instance. Table I shows the configuration parameters of c4i2p.

Finally, the `i2pProjectBBDD` machine runs a MySQL DataBase Management System (DBMS) to provide data persistence and storage for the rest of the crawling instances. The DBMS engine has to be configured to support a high number of concurrent connections. As it can be seen in Table I, every c4i2p instance simultaneously runs 50 discovery threads (the discoverers), the DBMS engine being configured to support a minimum of 10 (VMs) × 50 (threads) = 500 concurrent connections.

V. Results on I2P Eppsites and Discussion

The experiment took a total of 111 days, it starting on June 7th 2019 and finishing on September 26th 2019. During this period we managed a total of 54,974 I2P URLs, most of them obtained from leaseSets of floodfill routers. They correspond to services that can be web related (eepsites) or not. In fact, only 787 of them were eepsites and, thus, they were successfully crawled.

Figures 4 and 5 show the number of daily observed services during the experiment. As expected, the number of eepsites found (green dots in the figures) is much lower than the number of total observed services in the darknet (blue crosses in the figures). In Figure 5, a detailed view of the total number of eepsites is shown together with their cumulated sum (blue dots in the figure). Additionally, a SMA (Simple Moving Average) for 5 days (continuous orange line in both figures) is computed in order to see trends in the discovery and crawling processes.

As it can be seen in the previous figures, new I2P services are found every day. Indeed, a daily average of ∼ 500 services is computed at the end of the experiment. Similarly, eepsites are crawled almost every day, that resulting in an average of ∼ 5 which motivates the linear increment in the number of found eepsites from the beginning of the experiment shown in Figure 5 (blue dots). The high number of eepsites observed in the first day is motivated by the initial system contribution to the total number of services. These eepsites come from SEED, FLOODFILL and DISCOVERED sources. Moreover, no significant trends can apparently be seen for both total services and eepsites, according to the SMA values.

Some interesting conclusions can be also obtained from the analysis of the state and sources of the observed services. Figure 6 shows the rate of observed services in a specific state among those considered at the end of Section III-B3, ordered by their source. These results are numerically shown in Table II too. From them, it can be seen that the number of discarded eepsites is very high in comparison with the ones finally crawled (FINISHED). In fact, the number of FINISHED eepsites is only ∼ 1.5% (787) of the total observed services in the darknet. Notice that most of the services were obtained from FLOODFILL routers (50,784), the number of DISCOVERED ones being really low (292). This last value is clearly motivated by the low number of eepsites finally crawled. However, in percentage, the number of FINISHED eepsites obtained by crawling other eepsites (discovered) is higher than those obtained from FLOODFILL routers. Additionally, as expected, the higher number of DISCOVERED eepsites comes from FLOODFILL sources. Finally, 96.66% of services from SEED related sources are discarded. The initial SEED list was chosen manually from a previous work but, as the results show, it can be obsolete with most of the eepsites no longer available currently. As a consequence, we can conclude that the number of eepsites comprising the I2P network changes over time, some of them disappearing from the network while some others are created.

An additional comment should be done regarding the eepsites at the state DISCOVERING both in Figure 6 and Table II. Despite this is a transition state, its appearance as a final state in the results is due to the fact that our experimentation finished in time while some eepsites were still being contacted by the discovery procedure.

It is also interesting to see the distribution of the number of discovery attempts consumed for FINISHED eepsites. In Figure 7 the reader can see that in absolute (blue bars) and relative (orange bars) units. A deep inspection of the results concludes that only ∼ 20% of eepsites needed 1 attempt to be finally crawled. From that we can conclude the poor availability of eepsites in the I2P network. Another interesting result is the fact that some other nodes required an exceptional number of attempts to be finally crawled. Indeed, ∼ 13% of eepsites needed more than or equal to 200 attempts to be crawled. The main hypothesis behind this result is that these sites are continuously appearing and disappearing from the network and they are difficult to be reached with just few discovery attempts.

In summary, new I2P services are continuously appearing

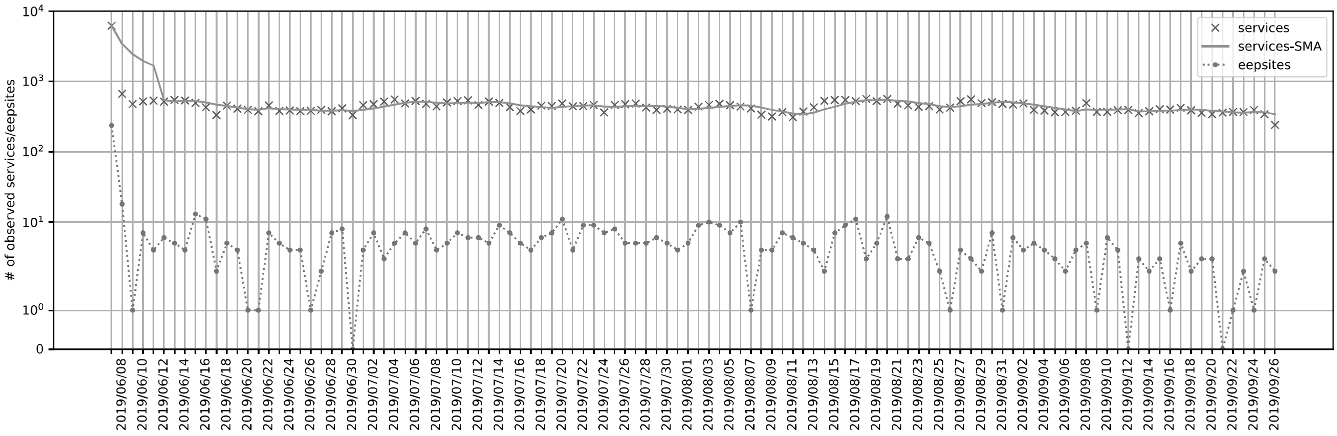


Figure 4. Daily number of observed I2P services (blue crosses), SMA (Simple Moving Average) (orange line) and eepsites (green dots). Notice the high number of services/eepsites at the beginning of the experiment where the system added more than 6,000 I2P services from initial seeds and floodfill sources.

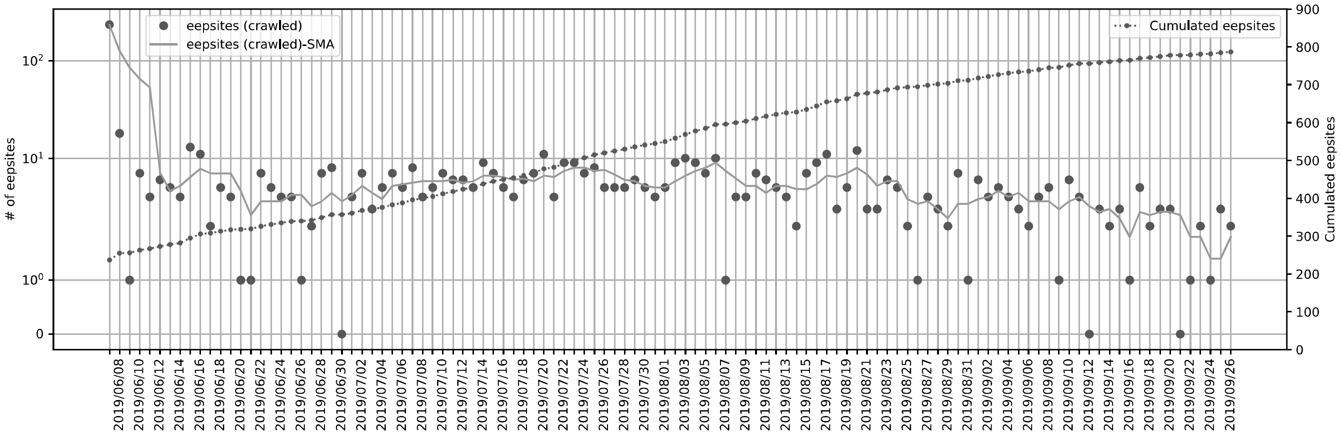


Figure 5. Detailed view of the daily rate of eepsites found and crawled (green dots), SMA (Simple Moving Average) (orange line) and cumulated sum (blue crosses). At the beginning of the experiment, the system crawled more than 200 eepsites from SEED, FLOODFILL and DISCOVERED sources.

Source	# services by source	State	# by state	% by state	# total services	% of total services
SEED	3,938	FINISHED	129	3.31	54,974	0.25
		DISCOVERING	0	0		0
		DISCARDED	3,768	96.66		6.85
FLOODFILL	50,784	FINISHED	600	1.18		1.09
		DISCOVERING	12,167	23.96		22.13
		DISCARDED	38,012	74.85		69.14
DISCOVERED	292	FINISHED	58	19.86		0.15
		DISCOVERING	3	1.03		0.005
		DISCARDED	230	78.77		0.42

Table II
SOURCE VS. STATE RESULTS: AGGREGATED RESULTS OBTAINED FROM THE EXPERIMENT DEPENDING ON THE SOURCE AND THE EEPSITES' STATE.

in the darknet though most of them are not eepsites and have been obtained from FLOODFILL routers. Moreover, most of active eepsites in the past, those included as SEED sources in this experiment, are not currently available. This could mean that, in general, eepsites have a short life. In fact, only 3.31% of the eepsites from SEED sources have been finally crawled according to Table II. As it will be seen in the next section, such eepsites correspond to either *source* or *sink* eepsites (nodes) in the darknet. Moreover, even having a low number of FINISHED eepsites from discovered sources, it is expected to find a higher number of eepsites obtained from crawling eepsites (discovered) than from other kind of sources due to their interconnections and relationships. Finally, active and persistent eepsites are easy to be discovered.

A. *Size analysis*

After the various results and behaviors previously described, we now perform an analysis about the size of I2P eepsites. For that, information about their number of pages is extracted during the crawling procedure. Additionally, the number of words, letters, images and scripts in the home page, as well as the predominant eepsite's language, are also estimated.

According to Figure 8, the I2P darknet is mostly composed

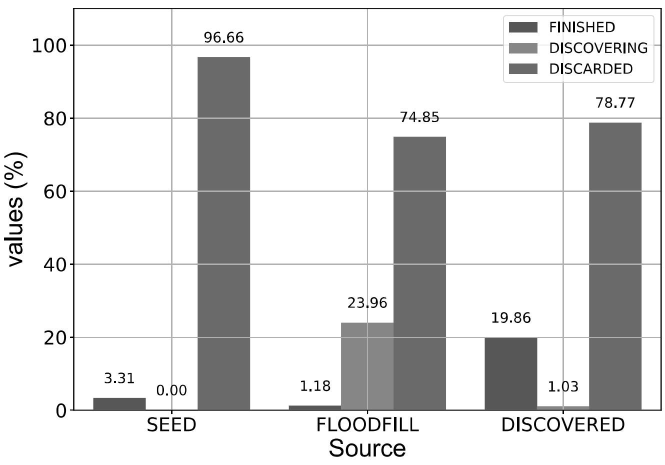


Figure 6. Source vs. state results: Percentage of services from a specific source in a specific state.

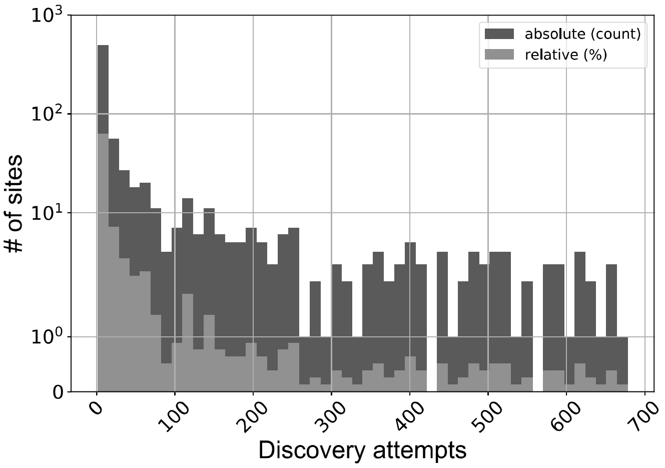


Figure 7. Distribution of discovery attempts for FINISHED eepsites.

of small eepsites. In fact, $\sim 80\%$ of the total number of eepsites have less than or equal to 30 pages. However, there are some exceptions. For instance, extraordinary large eepsites with 5, 6 or even 11 thousand pages exist, the largest one having 20,630 pages. Table III shows sites with more than 1,000 pages.

Regarding the size of eepsites related pages, Figure 9 shows that almost all eepsites have a home page with a few number of letters, words, images and scripts.

Finally, also an analysis about the predominant language used in eepsites is performed. To do that, we consider a widely accepted language detection engine: Google Translator. The results obtained are shown in Table IV, from which we can conclude that English is the most used language. However, we cannot be very confident about these results due to the small number of eepsites that are crawled.

B. Connectivity analysis

Several works in the literature have successfully addressed the study of the Surface Web connectivity as a graph [23], [24]. In this section, we will adopt a similar strategy to understand and unveil I2P eepsites relationships and organization patterns. First of all, a brief explanation of graph theory concepts is introduced. Second, specific results will be presented and discussed in detail.

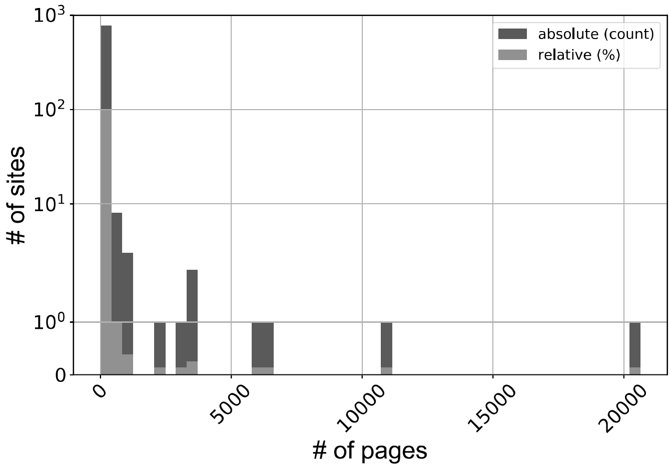


Figure 8. Eepsites size: Distribution of the number of pages of found eepsites.

Let $G = (V, E)$ be a directed graph comprising a set of nodes V (eepsites) and a set of edges E (connections). Each edge connects a pair of nodes (u, v) through a direct connection from u to v. This way, the number of edges from node u to some other nodes: $(u, v_1), (u, v_2), \cdots, (u, v_n)$, is denoted as *out-degree* (eepsite outgoing links). Similarly, the *in-degree* (eepsite incoming links) value of a node u corresponds to the number of edges pointing to node u from some other nodes: $(w_1, u), (w_2, u), \cdots, (w_m, u)$. Based on that, we can define four types of nodes: *source* nodes, with only outgoing links; *sink* nodes, with only incoming links; *connected* nodes, with both incoming and outgoing links; and *isolated* nodes, which are not connected to any other node in the graph.

The distribution of outgoing and incoming links is shown in Figures 10 and 11, respectively. We can observe that most of the crawled nodes have not significant out-degree or in-degree values. In fact, $\sim 10\%$ of the nodes correspond to source nodes, $\sim 12\%$ to sink nodes, and $\sim 12\%$ to connected nodes. Finally, $\sim 66\%$ of nodes are isolated nodes.

We should remark that there are some eepsites that can be considered as anomalous in terms of out- and in-degree values. Nodes with highest out- and in-degree values are summarized in Tables V and VI, respectively. In particular, the node with the highest value of outgoing links, 385 in total, is the `identiguy.i2p` eepsite. It contains a dynamically updated list of eepsites but does not cover the whole I2P, i.e. the hidden part. On the other hand, the node with the highest value of incoming links, 58 in total, corresponds to `forum.i2p`. It is worth noting that most of the biggest in-degree nodes were offline (discarded) probably due to two principal reasons: the number and relationships among the nodes are inconsistent, or many of the eepsites are unattended and not frequently updated.

In addition, all the highest out- and in-degree values correspond to eepsites found from SEED sources. Based on that, we can confirm the existence and persistence of some relevant I2P eepsites supporting a kind of network backbone among

# of pages	Eepsite
20630	ux6prousphswf56bym7yo7kst4ybh45y2z2wrnw7dujmrz56hq4q.b32.i2p
11059	qiii4iqrj3fwv4ucaj2oykcvsob75jviycv3ghw7dhzxg2kq53q.b32.i2p
6336	qa4boq364ndjdgow4kadycr5vvch7hofzblcqangh3nobzvyew7a.b32.i2p
5913	whba2ljn2sjvke45yjkyudzmelwkjcop3m7r6kubohngq3pb6cqa.b32.i2p
3614	a2lnpfsrhy5d3yky6xsut6gj6j76vn3lsy7kvabvedtu2d37s65q.b32.i2p
3516	gwqdodo2stgwgwusekxpkh3hbtph5jjc3kovmov2e2fbfdxg3woq.b32.i2p
3118	u6pciacxnpbsq7nwc3tgutywochfd6aysgayijr7jxzoysgxklvq.b32.i2p
2186	i2pforum.i2p
1194	25cb5kixhxm6i6c6wequrhi65mez4duc4l5qk6ictbik3tnxlu6a.b32.i2p

Table III
I2P EEPSITES WITH MORE THAN 1,000 PAGES.

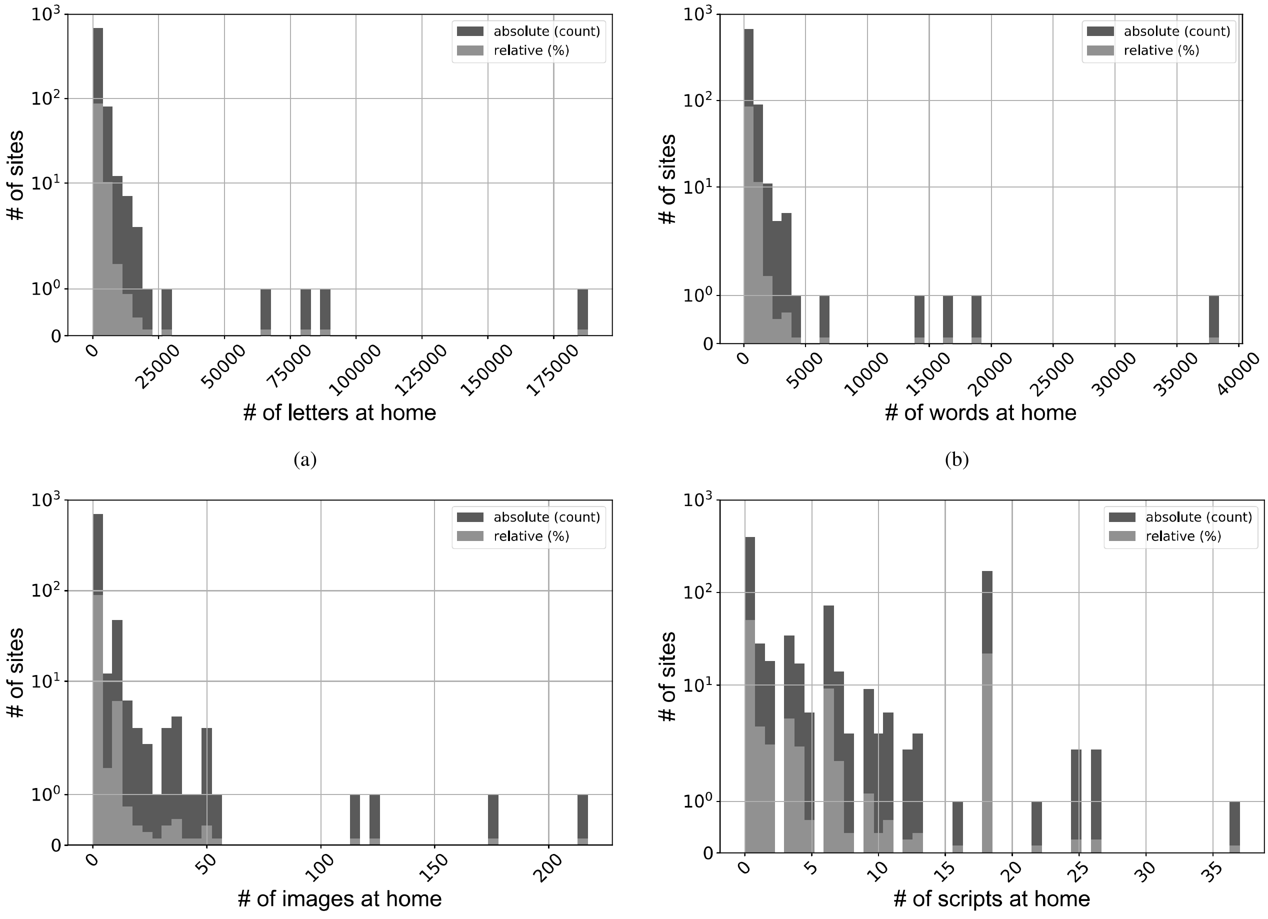


Figure 9. Eepsites content analysis in terms of number of (a) letters, (b) words, (c) images and (d) scripts.

Table IV
LANGUAGE OF EEPSITES ACCORDING TO GOOGLE TRANSLATOR ENGINE.

Language	%
English	96.31
French	0.86
German	0.86
Spanish	0.62
Norwegian	0.25
Latin	0.25
Italian	0.25
Welsh	0.12
Turkish	0.12
Portuguese	0.12
Dutch	0.12
Catalan	0.12

eepsites. However, there is also an important group of eepsites, the isolated ones. Moreover, with the aim of providing more information about these relevant eepsites, Table VII shows their specific contents.

Figures 12 and 13 depict the relationships among top eepsites with the highest out- and in-degree values. In the figures, the size of the labels and nodes is directly proportional to their out- and in-degree values, respectively, such values being depicted in brackets. As shown, almost all the nodes are either directly connected or the connection is carried out through an intermediate node.

An overall view of the connectivity of the I2P network is finally shown in Figure 14, where only nodes with at least one incoming or outgoing link are depicted. In this figure, the

out-degree	Eepsite	Status	Source
385	identiguy.i2p	FINISHED	SEED
248	i2pwiki.i2p	FINISHED	SEED
152	inr.i2p	FINISHED	SEED
70	s6lagaqbn572fvnr7vsxsqwbwxb6m3gr5hu6eetstykdm2opp2ia.b32.i2p	FINISHED	FLOODFILL
54	zy37tq6ynucp3ufoyeegswqjaeofmj57cpm5ecd7nbanh2h6f2ja.b32.i2p	FINISHED	SEED
45	5ypxuqf2ufqdsf3ejv5xwrgwatjxf2uw7tyxz2av44pka4w3pvza.b32.i2p	FINISHED	FLOODFILL
31	fex6v4zccrovs7dixqbigbbqtrb7ylrmpgphwnwoyjutrg56qmoa.b32.i2p	FINISHED	FLOODFILL
28	i2pforum.i2p	FINISHED	DISCOVERED
26	stats.i2p	FINISHED	DISCOVERED
26	pwgma3snbsgkddxgb54mrxxkt3l4jzchrtp52vxmw7rbkjygylxq.b32.i2p	FINISHED	SEED
19	trac.i2p2.i2p	FINISHED	SEED
18	i2p-projekt.i2p	FINISHED	SEED
17	isxls447iuumsb35pq5r3di6xrxr2igugvshqwhi5hj5gvhwvqba.b32.i2p	FINISHED	SEED
17	4bpcp4fmvyr46vb4kqjvtxlst6puz4r3dld24umooiy5mesxzspa.b32.i2p	FINISHED	SEED
15	uda2rkhskjdb4w7xiftz3btfpl7bhxsy5gwpiiiongte4gulbuza.b32.i2p	FINISHED	FLOODFILL

Table V
EEPSITES WITH LARGEST OUT-DEGREE VALUES.

in-degree	Eepsite	Status	Source
58	forum.i2p	DISCARDED	SEED
55	ugha.i2p	DISCARDED	SEED
52	www.i2p2.i2p	DISCARDED	SEED
49	i2p-projekt.i2p	FINISHED	SEED
47	no.i2p	DISCARDED	SEED
46	inproxy.tino.i2p	DISCARDED	SEED
45	i2host.i2p	DISCARDED	SEED
45	perv.i2p	DISCARDED	SEED
45	tino.i2p	DISCARDED	SEED
41	i2pwiki.i2p	FINISHED	SEED
21	sperrbezirk.i2p	DISCARDED	SEED
17	diftracker.i2p	FINISHED	SEED
13	visibility.i2p	DISCARDED	SEED
12	str4d.i2p	DISCARDED	SEED
11	www.imule.i2p	DISCARDED	SEED
11	bote.i2p	DISCARDED	SEED
11	killyourtv.i2p	DISCARDED	SEED

Table VI
EEPSITES WITH HIGHEST IN-DEGREE VALUES.

Eepsite	Content
identiguy.i2p	Centralized list of eepsites updated periodically
i2pwiki.i2p	I2P official Wiki
inr.i2p	Free domain register
s6lagaqbn572fvnr7vsxsqwbwxb6m3gr5hu6eetstykdm2opp2ia.b32.i2p	Private blog arguing about US politicians
zy37tq6ynucp3ufoyeegswqjaeofmj57cpm5ecd7nbanh2h6f2ja.b32.i2p	I2P tutorials
5ypxuqf2ufqdsf3ejv5xwrgwatjxf2uw7tyxz2av44pka4w3pvza.b32.i2p	DEFCON201 group forum in I2P (https://defcon201.org/)
fex6v4zccrovs7dixqbigbbqtrb7ylrmpgphwnwoyjutrg56qmoa.b32.i2p	Private forum
i2pforum.i2p	I2P support forum
stats.i2p	I2P statistics and NetDB
pwgma3snbsgkddxgb54mrxxkt3l4jzchrtp52vxmw7rbkjygylxq.b32.i2p	DEFCON201 group forum in I2P (https://defcon201.org/)
trac.i2p2.i2p	I2P bug tracker (https://trac.i2p2.de/)
i2p-projekt.i2p	Official I2P eepsite
isxls447iuumsb35pq5r3di6xrxr2igugvshqwhi5hj5gvhwvqba.b32.i2p	Private user eepsite. It seems to be built with the default I2P web template
4bpcp4fmvyr46vb4kqjvtxlst6puz4r3dld24umooiy5mesxzspa.b32.i2p	Private user eepsite. It seems to be built with the default I2P web template
uda2rkhskjdb4w7xiftz3btfpl7bhxsy5gwpiiiongte4gulbuza.b32.i2p	Private blog talking about several topics
diftracker.i2p	Torrent tracker

Table VII
TOP IN- AND OUT-DEGREE EEPSITES CONTENTS.

neighborhood for node identiguy.i2p can be observed, where all the nodes (colored nodes and links) can be accessed identiguy.i2p in three or less hops. It is also remarkable the existence of pairs of nodes separated from the rest. For the sake of clarity, only eepsites with highest out- and in-degree values are highlighted in the figure, though interested readers are kindly referred to the interactive version of this graph at https://nesg.ugr.es/i2p. Based on it, much more details of nodes and relationships can be obtained.

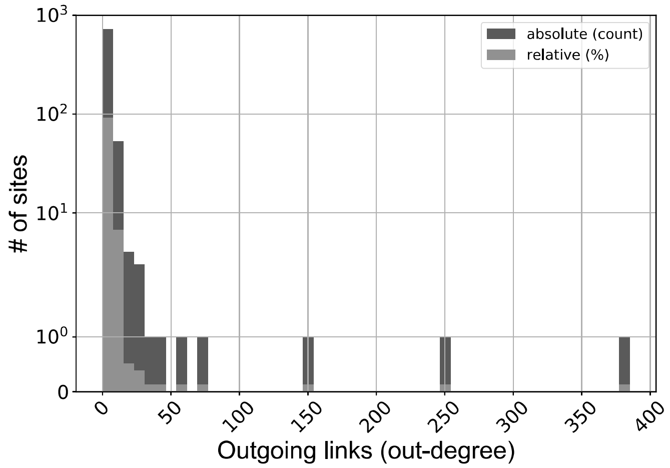


Figure 10. Distribution of eepsite outgoing links (out-degree).

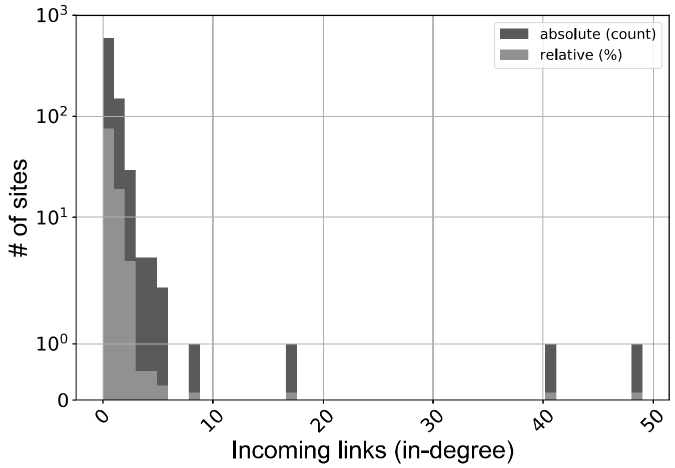
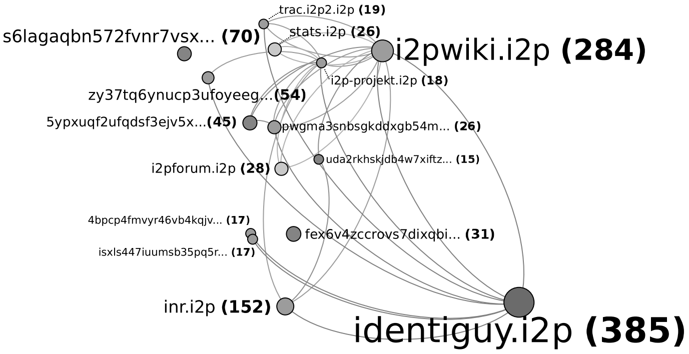


Figure 11. Distribution of eepsite incoming links (in-degree).



trac.i2p2.i2p (19)
s6lagaqbn572fvnr7vsx... (70)
stats.i2p (26)
i2pwiki.i2p (284)
i2p-projekt.i2p (18)
zy37tq6ynucp3ufoyeeg...(54)
5ypxuqf2ufqdsf3ejv5x...(45)
pwgma3snbsgkddxgb54m... (26)
uda2rkhskjdb4w7xiftz... (15)
i2pforum.i2p (28)
4bpcp4fmvyr46vb4kqjv... (17)
fex6v4zccrovs7dixqbi... (31)
isxls447iuumsb35pq5r... (17)
inr.i2p (152)
identiguy.i2p (385)

Figure 12. Connectivity graph for nodes with highest out-degree values.

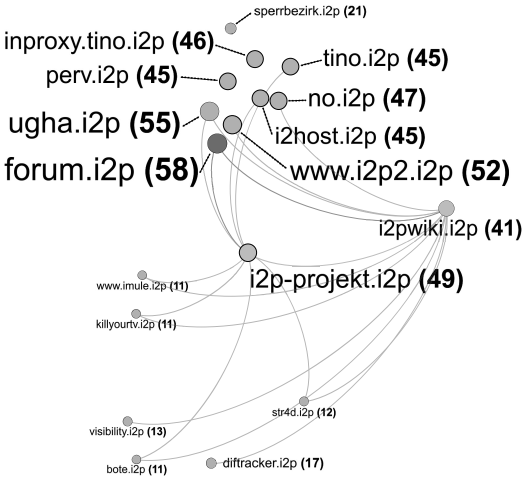


Figure 13. Connectivity graph for nodes with highest in-degree values.

VI. Conclusions and future work

Crawling the Surface Web has been widely addressed by the research community with different aims. However, the Deep Web part is still highly unknown despite its big interest not only for researchers, but also for authorities and security forces.

In this work, we inspect the I2P darknet by devising new methods and tools based on crawling procedures. The results obtained conclude that I2P is a highly decentralized network, where new services dynamically appear and disappear over time. A small portion of the I2P services are web related services, called eepsites. Regarding content and size aspects, it can be concluded that eepsites are mainly small websites. From the perspective of connectivity and relationships, eepsites are very disconnected from the rest, i.e., they present low in- and out-going connectivity degrees. On the contrary, few of them highlight from the rest in the same terms thus having a kind of eepsite network backbone. Moreover, more than a half of the total observed eepsites in our experimentation were completely isolated, thus becoming a hidden part of the overall set of eepsites.

Beyond the results obtained, some further work should be done. This way, we plan to extend the experimentation to cover as many I2P nodes as possible. Another relevant focus is to study isolated eepsites and those separated from the rest by making groups. It would be also interesting to analyze some other type of services used in I2P. Finally, aimed to get a more in-depth knowledge of the Deep Web, it would be necessary to extend the tool to address other darknet technologies, like Freenet (https://freenetproject.org/) for a further comparison.

Acknowledgement

This work has been partially supported by Spanish Government-MINECO (Ministerio de Economía y Competitividad), the ERDF (European Regional Development Fund) through project TIN2017-83494-R (MDSM) and the 5B starting research grants program of the University of Granada.

References

[1] M. K. Bergman, “White Paper: The Deep Web: Surfacing Hidden Value,” *Journal of Electronic Publishing*, vol. 7, no. 1, 2001.

[2] M. W. Al-Nabki, E. Fidalgo, E. Alegre, and L. Fernández-Robles, “TORank: Identifying the most influential suspicious domains in the tor network,” *Expert Systems with Applications*, vol. 123, pp. 212–226, 2019.

[3] G. Kalpakis, T. Tsikrika, N. Cunningham, C. Iliou, S. Vrochidis, J. Middleton, and I. Kompatsiaris, “OSINT and the Dark Web,” in *Open Source Intelligence Investigation: From Strategy to Implementation*, ser. Advanced Sciences and Technologies for Security Applications, B. Akhgar, P. S. Bayerl, and F. Sampson, Eds. Cham: Springer International Publishing, 2016, pp. 111–132.

[4] N. P. Hoang, P. Kintis, M. Antonakakis, and M. Polychronakis, “An empirical study of the I2P anonymity network and its censorship resistance,” in *Proceedings of the Internet Measurement Conference 2018*. ACM, 2018, pp. 379–392.

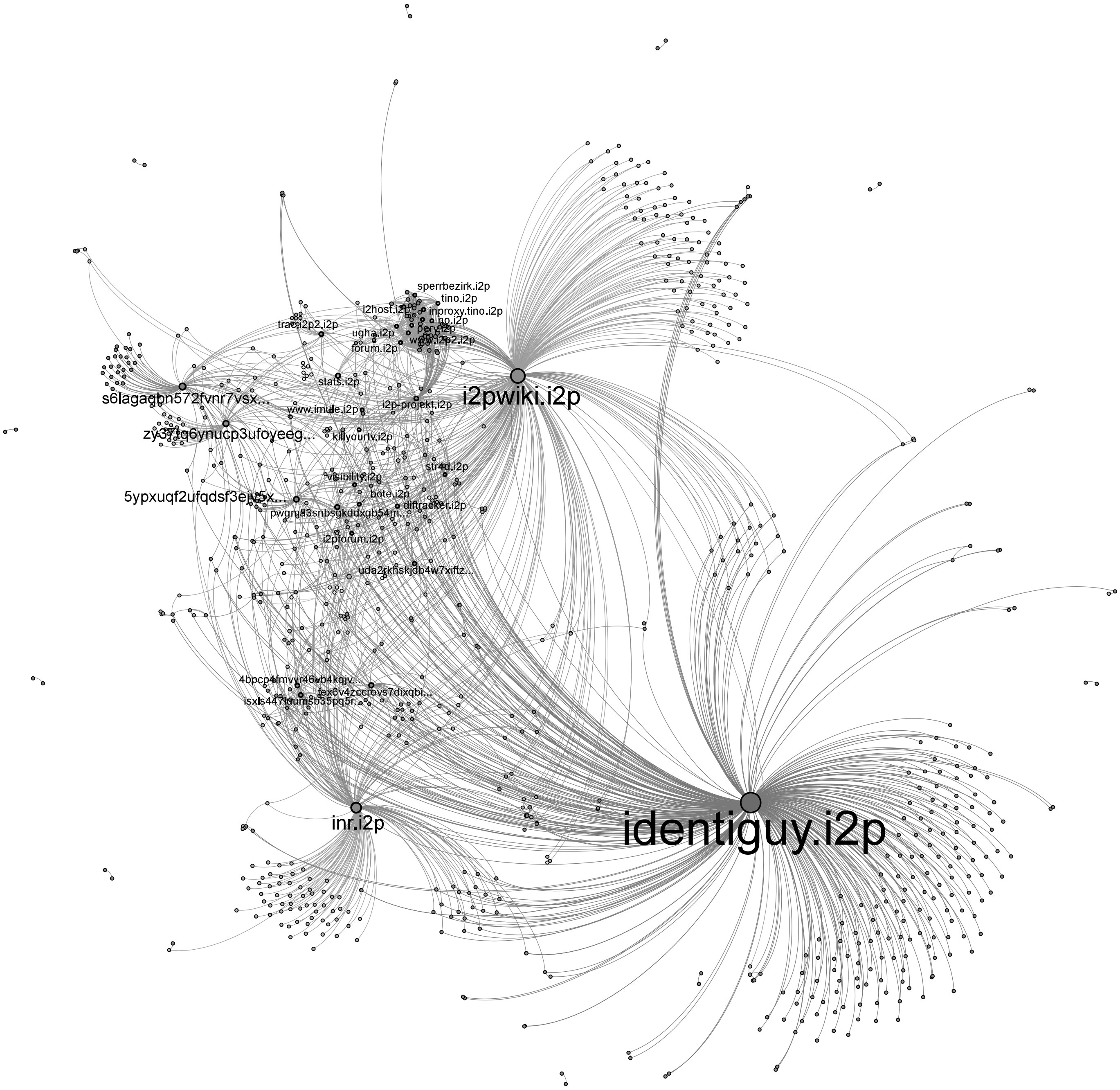


Figure 14. Overall view of the I2P eepsites with at least one incoming or outgoing link.

[5] A. Kumar and E. Rosenbach, "The Truth About the Dark Web," *Finance & Development*, vol. 56, no. 3, pp. 22–25, 2019. [Online]. Available: https://www.imf.org/external/pubs/ft/fandd/2019/09/pdf/the-truth-about-the-dark-web-kumar.pdf

[6] A. Biryukov, I. Pustogarov, F. Thill, and R.-P. Weinmann, "Content and popularity analysis of tor hidden services," in *2014 IEEE 34th International Conference on Distributed Computing Systems Workshops (ICDCSW)*. IEEE, 2014, pp. 188–193.

[7] M. W. Al Nabki, E. Fidalgo, E. Alegre, and I. de Paz, "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*, 2017, pp. 35–43.

[8] A. Biryukov, I. Pustogarov, and R.-P. Weinmann, "Trawling for tor hidden services: Detection, measurement, deanonymization," in *2013 IEEE Symposium on Security and Privacy*. IEEE, 2013, pp. 80–94.

[9] G. Kadianakis and K. Loesing, "Extrapolating network totals from hidden-service statistics," The Tor Project, Tech. Rep. 2015-01-001, 2015. [Online]. Available: https://research.torproject.org/techreports/extrapolating-hidserv-stats-2015-01-31.pdf

[10] A. Montieri, D. Ciuonzo, G. Bovenzi, V. Persico, and A. Pescapé, "A Dive into the Dark Web: Hierarchical Traffic Classification of Anonymity Tools," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 3, pp. 1043–1054, 2020.

[11] A. Montieri, D. Ciuonzo, G. Aceto, and A. Pescapé, "Anonymity Services Tor, I2P, JonDonym: Classifying in the Dark (Web)," *IEEE Transactions on Dependable and Secure Computing*, vol. 17, no. 3, pp. 662–675, 2020.

[12] Z. Cai, B. Jiang, Z. Lu, J. Liu, and P. Ma, "isAnon: Flow-Based Anonymity Network Traffic Identification Using Extreme Gradient Boosting," in *2019 International Joint Conference on Neural Networks (IJCNN)*, 2019, pp. 1–8.

[13] H. Yin and Y. He, "I2P Anonymous Traffic Detection and Identification," in *2019 5th International Conference on Advanced Computing Communication Systems (ICACCS)*, 2019, pp. 157–162.

[14] K. Shahbar and A. N. Zincir-Heywood, "How far can we push flow analysis to identify encrypted anonymity network traffic?" in *NOMS 2018 - 2018 IEEE/IFIP Network Operations and Management Sympo-*

sium, 2018, pp. 1–6.

[15] G. Avarikioti, R. Brunner, A. Kiayias, R. Wattenhofer, and D. Zindros, "Structure and content of the visible darknet," *arXiv preprint arXiv:1811.01348*, 2018.

[16] I. Sanchez-Rola, D. Balzarotti, and I. Santos, "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*. International World Wide Web Conferences Steering Committee, 2017, pp. 1251–1260.

[17] Y. Gao, Q. Tan, J. Shi, X. Wang, and M. Chen, "Large-scale discovery and empirical analysis for I2P eepsites," in *2017 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 2017, pp. 444–449.

[18] E. Spertus, "Parasite: Mining structural information on the web," *Computer Networks and ISDN Systems*, vol. 29, no. 8-13, pp. 1205–1215, 1997.

[19] S. J. Carrière and R. Kazman, "Webquery: Searching and visualizing the web through connectivity," *Computer Networks and ISDN Systems*, vol. 29, no. 8-13, pp. 1257–1267, 1997.

[20] S. Lawrence and C. L. Giles, "Accessibility of information on the web," *Nature*, vol. 400, no. 6740, p. 107, 1999.

[21] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer networks and ISDN systems*, vol. 30, no. 1-7, pp. 107–117, 1998.

[22] K. Bharat and M. R. Henzinger, "Improved algorithms for topic distillation in a hyperlinked environment," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '98. New York, NY, USA: Association for Computing Machinery, 1998, pp. 104–111.

[23] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins, "The web as a graph: measurements, models, and methods," in *International Computing and Combinatorics Conference*. Springer, 1999, pp. 1–17.

[24] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, "Graph structure in the web," *Computer networks*, vol. 33, no. 1-6, pp. 309–320, 2000.

[25] R. Meusel, S. Vigna, O. Lehmberg, and C. Bizer, "Graph structure in the web—revisited: a trick of the heavy tail," in *Proceedings of the 23rd international conference on World Wide Web*. ACM, 2014, pp. 427–432.

[26] R. Baeza-Yates and C. Castillo, "Relating web characteristics with link based web page ranking," in *Proceedings Eighth Symposium on String Processing and Information Retrieval*. IEEE, 2001, pp. 21–32.

[27] M. R. Henzinger, "Hyperlink analysis for the web," *IEEE Internet computing*, vol. 5, no. 1, pp. 45–50, 2001.

[28] C. H. Ding, H. Zha, X. He, P. Husbands, and H. D. Simon, "Link analysis: hubs and authorities on the world wide web," *SIAM review*, vol. 46, no. 2, pp. 256–268, 2004.

[29] C. Iliou, G. Kalpakis, T. Tsikrika, S. Vrochidis, and I. Kompatsiaris, "Hybrid focused crawling for homemade explosives discovery on surface and dark web," in *2016 11th International Conference on Availability, Reliability and Security (ARES)*. IEEE, 2016, pp. 229–234.

[30] J. P. Timpanaro, I. Chrisment, and O. Festor, "Monitoring the I2P network," *[Research Report] RR-7844, INRIA*, p. 16, 2011.

[31] R. A. Rodríguez-Gómez, G. Maciá-Fernández, and A. Casares-Andrés, "On understanding the existence of a deep torrent," *IEEE Communications Magazine*, vol. 55, no. 7, pp. 64–69, 2017.

[32] P. Liu, L. Wang, Q. Tan, Q. Li, X. Wang, and J. Shi, "Empirical measurement and analysis of I2P routers," *Journal of Networks*, vol. 9, no. 9, p. 2269, 2014.

[33] M. W. Al-Nabki, E. Fidalgo Fernández, E. Alegre Gutiérrez, V. González Castro *et al.*, "Detecting emerging products in TOR network based on k-shell graph decomposition," in *Proceeding of the 3th Spanish Conference on Cybersecurity Research*, 2017, pp. 24–30.

[34] L. Liu, H. Zhang, J. Shi, X. Yu, and H. Xu, "I2P anonymous communication network measurement and analysis," in *International Conference on Smart Computing and Communication*. Springer, 2019, pp. 105–115.

[35] T. Sochor, "Fuzzy control of configuration of web anonymization using TOR," in *2013 The International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE)*. IEEE, 2013, pp. 115–120.

[36] K. Shahbar and A. N. Zincir-Heywood, "Effects of shared bandwidth on anonymity of the I2P network users," in *2017 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2017, pp. 235–240.

[37] E. Erdin, C. Zachor, and M. H. Gunes, "How to find hidden users: A survey of attacks on anonymity networks," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 2296–2316, 2015.

[38] A. Ali, M. Khan, M. Saddique, U. Pirzada, M. Zohaib, I. Ahmad, and N. Debnath, "TOR vs I2P: A comparative study," in *2016 IEEE International Conference on Industrial Technology (ICIT)*. IEEE, 2016, pp. 1748–1751.

[39] J. P. Timpanaro, I. Chrisment, and O. Festor, "A bird's eye view on the I2P anonymous file-sharing environment," in *International Conference on Network and System Security*. Springer, 2012, pp. 135–148.

[40] M. Wilson and B. Bazli, "Forensic analysis of I2P activities," in *2016 22nd International Conference on Automation and Computing (ICAC)*. IEEE, 2016, pp. 529–534.

[41] J. P. Timpanaro, T. Cholez, I. Chrisment, and O. Festor, "Evaluation of the anonymous I2P network's design choices against performance and security," in *2015 International Conference on Information Systems Security and Privacy (ICISSP)*. IEEE, 2015, pp. 1–10.

[42] TOR, "TOR metrics project," https://metrics.torproject.org, 2020, [Online; Accessed 12-24-2020].

[43] I2P, "I2p official site," https://geti2p.net/, 2020, [Online; Accessed 12-11-2019].

[44] D. Echeverri Montoya, *Deep Web: TOR, FreeNET & I2P. Privacidad y Anonimato*. 0xWord, 2016.

[45] J. Nurmi, "Ahmia crawler official site," https://github.com/ahmia/ahmia-crawler, 2015, [Online; Accessed 12-24-2020].

[46] R. Magán-Carrión, A. Abellán-Galera, and G. Maciá-Fernández, "C4i2p official repository code," https://github.com/nesg-ugr/I2P_Crawler, 2019, [Online; accessed 12-24-2020].

[47] M. T. Khan, J. DeBlasio, G. M. Voelker, A. C. Snoeren, C. Kanich, and N. Vallina-Rodriguez, "An Empirical Analysis of the Commercial VPN Ecosystem," in *Proceedings of the Internet Measurement Conference 2018*, ser. IMC '18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 443–456.