Cutting Onions With Others' Hands: A First Measurement of Tor Proxies in the Wild

Dongqi Han*[‡], Shangdong Wang*[‡], Zhize He[¶], Zhiliang Wang *[‡], Wenqi Chen*[‡], Chenglong Li*[‡],
Jiahai Yang*[‡][§], Xingang Shi*[‡] and Xia Yin[†][‡]
*Institute for Network Sciences and Cyberspace, BNRist, Tsinghua University, Beijing, China
[†]Department of Computer Science and Technology, BNRist, Tsinghua University, Beijing, China
[‡]Zhongguancun Laboratory, Beijing, China
[§]Quan Cheng Laboratory, Jinan, Shandong, China
[¶]QI-ANXIN Technology Research Institute, Beijing, China

***Abstract*—Tor Proxies are principally designed for allowing users to easily access hidden services in Tor network through standard browsers instead of dedicated software (e.g., Tor Browser), which is user-friendly but at the expense of anonymity. At present, Tor proxies are volunteer-run without any central management. As a result, it still remains unclear the scale and usage of Tor proxies in the wild throughout the Internet. In light of this, we perform the first large-scale measurement study of Tor proxies through passively identifying Tor proxies worldwide and then conduct initial analysis based on the identified Tor proxies. We cooperate with one of the most popular public DNS resolvers and collect passive DNS for the last 4 years with hundreds of billions of raw DNS records per day. To analyze the usage of Tor proxies, we also crawl newly observed webpages of hidden services accessed via Tor proxies for two years. We propose several techniques to identify the Tor proxies and find about 700 historically valid and 130 online Tor proxies in over 30 countries. We also provide several insightful findings and promising directions to motivate future work on this topic.**

***Index Terms*—Network measurement, Tor proxies, Tor, hidden services**

I. Introduction

Nowadays, Tor [1] holds arguably the largest deployed anonymity overlay network with thousands of voluntary relays and millions of users from all around the world [2]. Tor plays an important role in maintaining users' anonymity through privacy-enhancing technology. In addition to user anonymity, Tor provides *hidden/onion services* for the anonymity of content publishers (i.e., servers). Onion services are configured to be accessed only through their own domain address with a special-use top-level domain (TLD) "`.onion`" which is only reachable via Tor network [3]. Therefore, accessing onion services via Tor certainly enables the anonymity of both users and servers.

However, the traces of accessing onion services have not disappeared thoroughly on the surface of the Internet. There is a *Tor proxy* mechanism to allow users to easily access onion services through proxies instead of Tor-dedicated software such as Tor Browser. In a nutshell, Tor proxies receive requests from users, and then access the services requested by the users through Tor network, and finally forward the response to the users. Note that the entire process is transparent for users. They can simply add (or change) the domain suffix of onion services. For example, if a user wants to access `service.onion`. He/She can simply visit `service.onion.proxy` with any favorite standard browser such as Google Chrome. The proxy with the domain name of `onion.proxy` will help the user complete the underlying request and content forwarding.

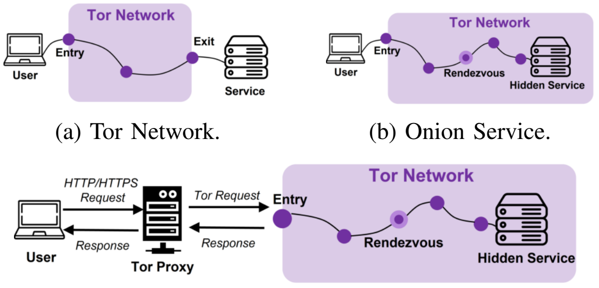
At present, Tor proxies are volunteer-run on self-configured servers [4], [5] without central management. As a result, it still remains unclear the *scale* (e.g., which are and how many) and *usage* (e.g., organizers and accessed services) of Tor proxies throughout the Internet. Besides, there has been an officially suggested security warning of Tor proxies to encourage users to the Tor Browser for real anonymity [4]. It is also worth studying why and to what extent users/proxies sacrifice/violate anonymity when using Tor proxies.

To answer the above questions, we resort to passive DNS data from one of the most popular public DNS service providers in China, which generates hundreds of billions of DNS records each day. We collect passive DNS data for four years from March 2018 to March 2022. (1) To measure Tor proxies throughout the Internet, we present several skills to identify valid onion service addresses and proxies from hundreds of billions of entries, followed by providing statistics and analysis of their current scale and status (in §III). (2) Based on the identified Tor proxies, we crawl the webpages of newly observed onion service addresses accessed via the proxies in the last two years. We then analyze the usage of Tor proxies from the *user side* through service classification to observe the preference as well as the *proxy side* through observing content modification to reveal abnormal behaviors of Tor proxies (in §IV).

Contributions. The contributions of this study include:

- We perform the first large-scale measurement study to passively identify Tor proxies worldwide. Our study is based on passive DNS data collected from one of the most popular public DNS resolvers worldwide with hundreds of billions of raw DNS records each day in 4 years from 2018 to 2022.

ISBN 978-3-903176-57-7© 2023 IFIP



(a) Tor Network.

(b) Onion Service.

(c) Tor Proxy.

Fig. 1: Tor network, onion services, and Tor proxies.

- We propose techniques to identify and validate Tor proxies from passive DNS records. In total, we find $\sim$700 historically valid Tor proxies in over 30 countries, of which more than 130 are online at measurement time.
- We conduct several analyses based on the identified Tor proxies. We analyze the usage of Tor proxies by crawling newly observed webpages of onion services requested via proxies for the last two years. We conduct experiments to reveal the usage of Tor proxies from the user side and content modification from the proxy side.

II. Background and Related Work

Tor. The Onion Router (Tor) [1] is an open-source software for building a worldwide anonymous overlay network throughout the Internet, consisting of more than 6,000 relays and millions of daily clients [6]. Tor provides anonymity for users by random communication and multiple encryptions through a network consisting of volunteer-operate relays. As shown in Fig. 1a, only the relays through the selected path can correctly decrypt and forward the requests. As a result, only the entry relay knows the information of the user and only the exit relay knows the connecting service, thereby providing users with anonymity.

Onion Services. In addition to user anonymity, Tor can also provide anonymity to content publishers (i.e., services) by introducing *onion services* or *hidden services*[1]. With a special-use top-level domain (TLD) "`.onion`" in their domain address, onion services can only be accessed through Tor network [3]. As shown in Fig. 1b, the user and server both build their own circuit and communicate through a rendezvous relay. The entire communication is within the Tor network. The prefix of the onion service address is an elaborate string containing the public key of an individual onion service, allowing the self-authentication of the onion service (i.e., the access can only be responded to by the genuine service). There are two widely-used versions of onion service addresses: V2 and V3 [2]. V2 address is made up of a 16-character Base32 hash of the public key of a certain onion service, followed by "`.onion`", which was deprecated by the official project in Oct. 2021. V3 address enhances the security by introducing more items in the prefix to form a 56-character Base32 hash as follows:

$$\texttt{V3\_Addr} = \texttt{Base32(PUBKEY|CHECKSUM|VERSION)},$$

where $\texttt{CHECKSUM} = \texttt{Hash(string|PUBKEY|VERSION)[:2]}$.

Tor Proxies. Normally, users can only access onion services in Tor Network through dedicated software (to select relays and build circuits), such as Tor Browser. *Tor proxies* allow users to easily access onion services through proxies instead of Tor-dedicated software. As shown in Fig. 1c, Tor proxies receive the requests from users, and then access the services requested by the users through Tor network, and finally forward the response to the users. A Tor proxy service with a distinct registered domain suffix, e.g., `onion.ly`. When a user wants to access an onion service `abc.onion`, he/she can directly use Google Chrome (a standard web browser) to access `abc.onion.ly` and let the proxy (i.e., `onion.ly`) help to complete the request forwarding. Tor proxies are volunteer-run with self-configured servers. Tor2web [4], [5] is a famous open-source project providing the underlying configuration for Tor proxy servers. The implementation is similar to HTTP proxies, except that the proxy servers need to access the Tor network after receiving users' requests.

Related Work. There are several works providing measurements on the Tor network, such as estimation of bandwidth [7] and V2/V3 onion services [2], [8], and service performance evaluation and improvement [9], [10]. There are also many works on the measurements of onion services, including domain ranking [11], measurement on certain services [12], service discovery [13], [14] and service classification [14]–[18]. Another type of related work is the security analysis of Tor Network, including the security risks of onion routing [19], [20], inference of service to break the anonymity [21]–[26], and privacy analysis of onion services [27]. These studies are orthogonal since our scope in this study is Tor proxy instead of Tor network.

As for studies related to Tor proxy, existing ones mainly use Tor proxy as a tool. For example, resilient botnet command and control via Tor2Web is proposed in [28]. In [29], a honeypot is developed with scripts reaching the application via Tor proxies. Other works [30]–[32] treat Tor proxies as the crawlers in the surface web without visiting Tor network. To the best of our knowledge, we are the first to conduct the study of measurement and analysis of Tor proxies throughout the Internet.

III. Identifying Tor Proxies in the Wild

In this section, we introduce the method and results of measuring Tor proxies worldwide, including the method of data collection (§III-A), identification and validation of Tor proxies (§III-B), as well as the results and analysis (§III-C).

A. Data Collection

As mentioned before, there is no community providing unified management of Tor proxies on the Internet, and only

[1] Hidden service is a broader concept than onion service. In the following, we use A because. We use the term "onion service" as we focus on Tor network instead of other anonymity networks such as I2P.

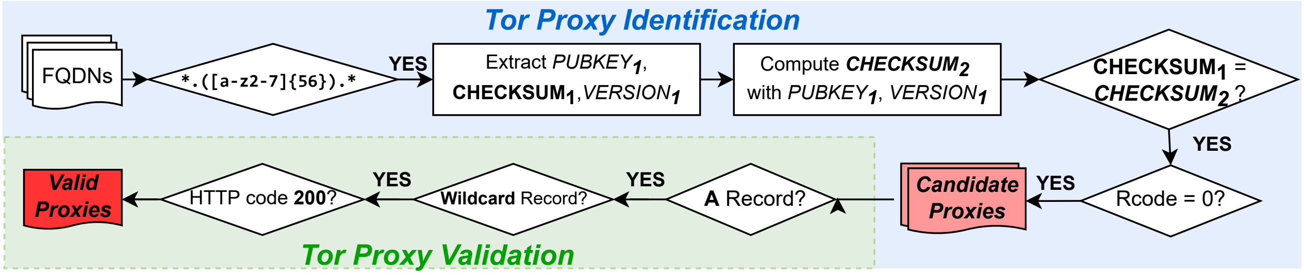


Fig. 2: The workflow of identifying and validating Tor proxies.

a dozen of well-known proxies can be found through search engines, which motivates us to build a more accurate, realistic, and broader picture of the Tor proxies in the wild. We have cooperated with the 114DNS provider, which is one of the largest DNS service providers in China [33]. The resolver IP addresses are 114.114.114.114 and 114.114.115.115.

We collect Passive DNS (PDNS) data for the last four years from March 2018 to March 2022[2] in a privacy-preserving manner. 114DNS resolvers have hundreds of billions of raw DNS records each day. In order to deal with such a huge amount of data, we *merge* the same request/response records by day (i.e., records with the same response that appear in one day are merged as one record with a total request number). Note that, to avoid raising ethical concerns, we only collect DNS *response* records consisting of FQDN (Fully Qualified Domain Name) and resolved responses without any user information (such as user IP) in this measurement. We also merge them *globally* over four years and finally get 53.4 billion distinct records in total.

B. Identification and Validation Method

Definition of Candidate Proxies and Valid Proxies. Before introducing our method, we define two kinds of Tor proxies:

- **Candidate Proxies**: Proxies that have been able to successfully resolve onion service addresses in the past observed from historical PDNS data.
- **Valid Proxies**: Proxies that can still respond to onion service requests at the time of measurement. Note that, *valid proxies* are a subset of *candidate proxies*.

The workflow of identifying *candidate proxies* and *valid proxies* is shown in Fig. 2. We introduce the two parts as follows.

1) PDNS Records → Candidate Proxies: It is arguably challenging to find the domain address of Tor proxies from a large number of PDNS records, which is tantamount to a needle in a haystack. Our high-level idea of identifying Tor proxies is to identify onion service addresses from all prefixes of FQDNs. As mentioned in §II, V3 onion service addresses use a checksum to ensure the integrity of the public key in the address, which can also be used to identify valid onion service addresses from all FQDNs. As shown on the top of Fig. 2, we first match FQDNs with the following regular expression:

⋆.([a-z2-7]{56}).⋆

which means the 56-character Base32 encoded strings as Base32 is made up of any letter of the alphabet, and decimal digits from 2 to 7. Here "⋆" means any string and "." is the separator of domains. In other words, we match the domain *label* (i.e., a part of a domain name separated by dots) with the characteristics of V3 address (recall V3 format in §II). We call ⋆.([a-z2-7]{56}) as the *prefix* of FDQN that consists of probable legitimate V3 address and its subdomain (the left ⋆), and the right ⋆ as the *suffix* of FDQN.

Next, we extract the public key and checksum field from the addresses and compare the checksum computed from extracted public key with the extracted checksum. A prefix of FQDN is a *legitimate* V3 address if two checksum fields are identical. In this case, if this FQDN record of a certain onion service has been successfully resolved (i.e., Rcode of the DNS response is 0), its suffix is considered as a candidate proxy. We introduce a configurable threshold K that represents the number of distinct legitimate V3 services verified by checksum. That is, we count the number of candidate proxies that were accessed by more than K legitimate V3 services. The purpose of K is to filter out some unpopular proxies (i.e., proxies that have only serviced for no more than K different services). The selection of K is empirically studied in §III-C.

2) Candidate Proxies → Valid Proxies: To validate the identified candidate proxies, we explore three characteristics of valid Tor proxies from their domain name resolution and web access response. We conduct a progressive validation of candidate proxies as shown in the bottom of Fig. 2. We first check whether the DNS resolution result of the proxy (domain suffix) has *A* record (whether within the valid period of registration) and *wildcard* record. A wildcard record is specified with a ⋆ as the leftmost label of a domain name (e.g., ⋆.onion.ly) and is typically used to handle unexpected domain requests. Tor proxy must configure the wildcard DNS resolution to take over any prefix onion services. The aforementioned two measurements from resolution are to check whether the domain retains the characteristics of valid Tor proxies. And finally, we directly access the webpage with an

[2] Due to data compliance and availability considerations, we are unable to obtain the latest data (i.e., the data from May 2022 to the present). Nonetheless, we believe that analyzing the trends over the past 4 years is sufficient to draw convincing conclusions (see §III-C), so the timeliness of the data has negligible impact on our work.

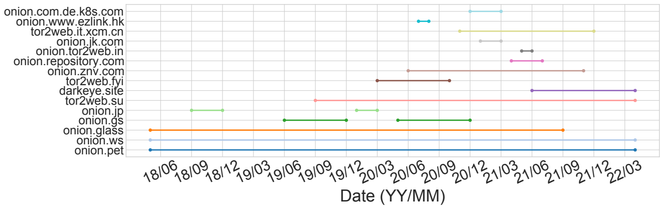


Fig. 3: The lifespan of representative Tor proxies.

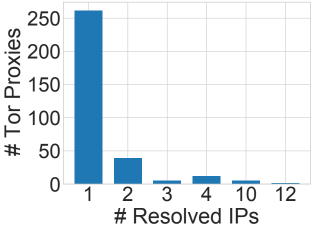


Fig. 4: The number of resolved IPs of proxies.

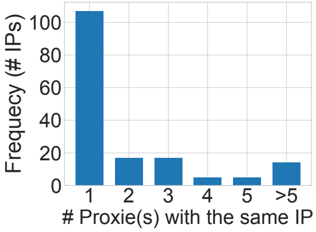


Fig. 5: Tor proxies with the same resolved IP.

active onion service followed by the proxy domain name and check the HTTP response status code. The reason we do not consider webpage content here is that we find most valid proxies will modify service content (see §IV-C). Although the HTTP response is the most authoritative way to identify valid Tor proxies, we still utilize the former two methods for observing candidate but not valid proxies (i.e. do not have valid responses but retain characteristics of Tor proxies).

C. Results and Analysis of Identified Proxies

We show the results of identified Tor proxies through the aforementioned measurements from the following aspects:

1) Scale of Tor Proxies: The results of identified proxies under different K are shown in Table I. As mentioned that K is an indicator of Tor proxy popularity, and we can observe that the number of proxies tends to stabilize when $K \geq 5$. Overall, the number of identified Tor proxies is considerable (nearly 700 when $K \geq 5$), compared with search engines that can only find a dozen of popular Tor proxies. Besides, we can also find that there are a large number of proxies that only serve one onion service (compare the number of candidate proxies when K=1 and $K > 1$), and most candidate proxies (e.g., $1-$ 133/697 $\approx$ 81% for $K = 5$) are not valid proxies (i.e., currently offline). Note that, the validation in Table I was conducted in May 2022, while the passive data began more than four years ago. This demonstrates the poor stability and reliability of most Tor proxies as they are volunteer-operated (e.g., for personal use or initial experiments). This may also explain the huge additional number of candidate proxies when $K = 1$. In the following experiments, we use 697 candidate proxies (i.e., $K = 5$) unless otherwise specified.

2) Lifespan: From Table I, we can also observe that quite a few candidate but not valid proxies still have A or wildcard resolution records, which means that they may be historically valid but expired now. We also conduct the long-term measurement of valid Tor proxies through multiple checks during

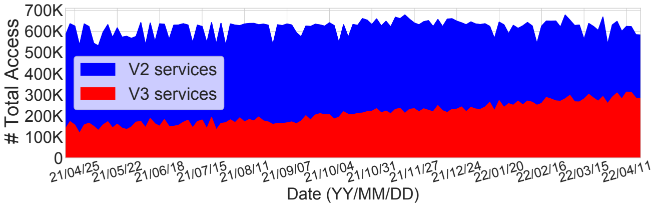


Fig. 7: The total number of onion service requests via Tor proxies by day.

one year. The results of three checks at intervals of about half a year are shown in Table II. The results witness the decrease of candidate proxies over time, especially those invalid ones with A/wildcard records.

To investigate the lifespan of individual proxies, we conduct another measurement here. Limited by the beginning time of this study, we are unable to validate proxies in real-time for four years. Accordingly, we propose an alternative method by observing whether there are any legitimate response records for each month. As this method may misjudge proxies with a small accessing scale, we choose representative proxies with different length of domain name according to the total number of requests. As shown in Fig. 3, each row represents a proxy, and the solid line plot indicates that the proxy is valid at the corresponding time, while the blank indicates that it is offline. We can find that only a small number of proxies (e.g., `onion.ws`, `onion.pet`, `tor2web.su`) can provide stable service. Some proxies (e.g., `onion.gs` and `onion.jp`) have long outages. Future work can conduct long-term active measurements to obtain more accurate results.

3) Resolved IPs: We resolve the domain names of identified Tor proxies to IPs (via dig or nslookup). In Fig. 4, we list the number of resolved IPs of proxies. Most of them have single or double IPs. We also investigate how many of these proxies are actually using the same IP, indicating that different proxies might have operators in common. As shown in Fig. 5, roughly 35% of IPs have more than one proxy running on them. We also observe the geographical distribution of proxies and the results of the top five countries with the most proxies are shown in Table III. The list of all countries is left in Appendix A. In short, We identify valid proxies in 31 countries, and the US takes up the majority of identified proxies and especially valid proxies across the world.

4) Domain Names: We analyze the label of domain names used by identified Tor proxies. Table IV lists some most frequently occurring domain labels and their frequencies. We can observe that frequently used domain labels are with a strong meaning of Tor proxies, such as "onion", "tor2web", "darkeye", for better promotion and memorization. We also compute the maximum level (length) of these domain names used by proxies, and the results are shown in Fig. 6. Although SLDs (second-level domain) such as `onion.ly` are easier to memorize, we can observe that most of them use domain names with a length of three or four. We speculate the reason is that the price of SLD registration is too expensive for

TABLE I: Number of identified Tor proxies.

# Distinct V3 services	# Candidate proxies	# Has A/wildcard record	# Valid proxies
$K = 1$	1136	347/263	133
$K = 3$	714	301/248	133
$K = 5$	697	287/246	133
$K = 10$	689	281/244	133
$K = 50$	687	281/244	133

TABLE II: Multiple validation checks in one year.

# Measurement time	# Has A/wildcard record	# Valid proxies
June 2021	378/268	138
Nov. 2021	333/281	135
May 2022	281/244	133

TABLE III: Top-five countries with most proxies.

Country Name (TOP 5)	# Candidate Proxies	# Valid Proxies
United States of America	200	82
China	98	14
Germany	30	12
Virgin Islands (British)	21	1
Singapore	14	2

TABLE IV: Frequently used domain labels of proxies.

Domain	Frequency
onion	552
tor2web	159
darkeye	30
darktor	16
hiddenservice	9
d2web	9

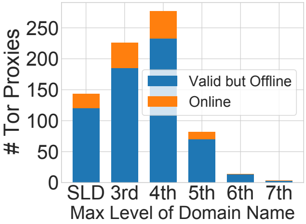


Fig. 6: The maximum level of the domain name of proxies.

volunteers. For example, the price of `onion.org` is up to $56,000 a year in NameSilo [34], a famous domain name registrar company.

5) Access Interests: We observe the access scale of onion services via Tor proxies from the number of DNS requests/responses in PDNS records. Note that, DNS requests do not necessarily mean active visits and the relationship between requests and accesses is not strictly 1:1 due to DNS cache. However, requests can reflect the access interests and rough scale. We leave the accurate measurement or estimation for future work. The result of requesting via identified proxies by day is shown in Fig. 7. We show the number of daily requests for V2 and V3 services separately. Here we identify V2 onion service addresses by matching FQDNs with the regular expression of "`*.([a-z2-7]{16}).[identified_proxies]`", where "`[identified_proxies]`" is the list of identified Tor proxy names. One may argue that some *noisy* domains with a string of 16-chars of "`[a-z2-7]`" are also coincidentally considered here as V2 addresses do not include a checksum. In order to observe such possible "false positive", we sample 10k requests and find that more than 95% of the addresses are known valid onion service addresses. The ratio of confirmed false positives (such as "`aaaaaaaabbbbbbbb`" and with obvious semantics like English words, Chinese Pinyin, etc.) is less than 1%. This observation indicates that the error on the V2 services in Fig. 7 is insignificant.

As shown in the results, the number of daily requests remains approximately 0.6M. The relatively stable number demonstrates a non-negligible phenomenon of using Tor proxies to access onion services. We can also observe that V2 services make up the majority but the percentage of V3 services increases over time. This is because Tor proxies still support V2 onion services, but the official Tor client does not anymore. In Appendix B, we also measure the request number by each proxy and witness the long-tailed distribution (Fig. 11). The brief conclusion is that the accessing is long-tailed, with a few well-known proxies taking up most of the accesses.

IV. Usage of Identified Tor Proxies

In this section, we analyze the usage of Tor proxies from *user* and *proxy* side by classifying onion services requested via identified proxies (§IV-B) and measuring the difference between webpages from proxies and the original ones (§IV-C).

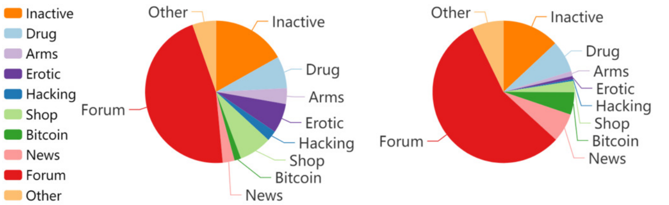
A. Data Collection

We collect the webpage of onion services that have been requested via Tor proxies to make a classification. As mentioned before, the lifespan of many onion services is short, which makes the collection difficult. Therefore, we conduct a long-term measurement of crawling the webpage of *newly observed* services (i.e. prefix of FQDN) each day for the last 2 years. Specifically, we observe the *unseen* FQDNs (appeared for the first time) every day, and only keep those prefixes/services with the suffixes of identified Tor proxies. To avoid illegal materials such as videos or pictures of child pornography or drugs, we only use the textual content, namely HTML code of the webpage of services [18]. Finally, we get 4,805 webpages of onion services accessed by Tor proxies in total.

B. Which classes of services are preferred?

Service Classification. We classify the collected webpages (in §IV-A) through Machine Learning and manual correction. First, we extract vectors consisting of weighting factors of keywords in crawled textual contents via TF-IDF [35], a common method to extract well-structured features from structure-free documents by calculating the term and inverse document frequency. Due to the lack of trusted labeled onion services, we leverage an *unsupervised* method K-Means to separate the webpages into clusters. Based on the taxonomy of onion services in existing studies [14]–[18] and the clustering result of our measurements, we separate the crawled services into ten classes such as drug, erotic and arms. Their name and detailed description are in Table VII in Appendix C. To further improve the accuracy, we manually add some *keywords* to correct the unreliable cluster. Some selected keywords for each class are listed in Table VIII in Appendix C.

Results and Analysis. As shown in Fig. 8, we count the number of each class of services and the total number of requests in PDNS records. The top-3 classes (apart from "inactive" and "other") w.r.t. the number of unique services are Forum (59%), Shop (9.6%), and Drug (9.3%), while Forum (70%), Drug



(a) Number of Unique Services **(b) Number of Total Requests**

Order	Class	% The number of services (except inactive and other)	% The number of requests (except inactive and other)
1	Forum	59.33%	70.27%
2	Drug	9.36%	9.10%
3	News	3.48%	8.22%
4	Bitcoin	1.94%	6.40%
5	Shop	9.63%	3.21%
6	Arms	4.62%	1.42%
7	Erotic	8.63%	0.92%
8	Hacking	3.01%	0.46%

Fig. 8: The proportion of each class of onion services accessed by Tor proxies.

(9.1%), and News (8.2%) are top-3 most frequently requested classes. The possible reason why users tend to visit forums and news could be less privacy-sensitive compared with other services (not absolutely). However, there is a notable ratio (9.1%) of accessing “Drug” services via proxies since it is illegal in most countries/regions. We suggest future work can further investigate to which extent the anonymity of users is compromised when accessing certain class of services.

Relationship between Proxies and Services. We also investigate the relationship between proxies and requested onion services. Fig. 9 depicts the relationship of proxies and classified onion services via proxies in an intuitive graph, where nodes are services or proxies and links are the requests of services via the proxies. We can also observe that many onion services are only served by one certain proxy. Moreover, we can find that users tend to request the same or the same class of services when using some proxies (see the edge of the figure). A possible reason is some onion services may also build proxies to promote their onion services to Internet users. This speculation could be used to analyze the relationship between onion services and proxies and between different proxies. However, to avoid raising ethical concerns, we do not gather any user information, thus it is difficult and beyond our scope to verify this speculation. Future work could conduct more comprehensive measurements and analyses.

C. Will Tor proxies modify the content?

Motivated by prior work on measuring content modification of HTTP proxies [36], we compare webpages for onion services requested by proxies with original ones via Tor browser. Thus, we additionally crawl the newly observed onion services through Tor network and compare them with the crawled webpages as mentioned in §IV-B. Considering that the modification may depend on the specific service, we compare the difference of webpages for each class of services, and divide the modification behavior into five categories, including “Expose” (exposing user access records to third-party tools such as Google Analytics), “Ad” (injecting ads not in the original page), “Redirect” (redirect to another webpage), “No Response” (no response for certain services), and “Other” (other content modification).

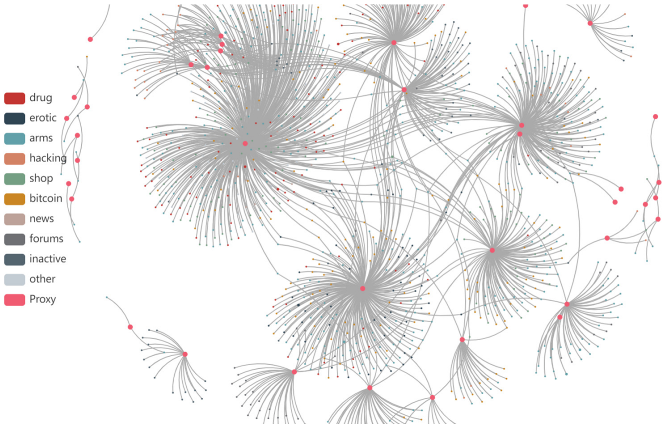


Fig. 9: The relationship of proxies and requested services.

TABLE V: Content modification of Tor proxies.

Class	# Unmodified	Expose	Ad	Redirect	No Response	Others	% Total Abnormal
Drug	4	28	33	44	18	8	97.04%
Erotic	3	28	31	59	9	5	97.78%
Arms	3	27	39	50	6	10	97.78%
Hacking	4	28	41	42	9	11	97.04%
Shop	3	27	31	43	9	22	97.78%
News	2	28	21	18	17	38	98.51%
Forum	1	27	46	38	9	14	99.25%
Bitcoin	3	27	41	43	9	12	97.78%
Total (Avg.)	**3**	**28**	**35**	**42**	**11**	**15**	**97.87%**

The results of the comparison are shown in Table V. We surprisingly find that almost all Tor proxies (∼98%) modify content (the rate for open HTTP proxies is only 5% according to [36]). Only one of over 130 Tor proxies is unmodified considering all classes of service. The top two types of modification are “Redirect” and “Ad”. Appendix C lists an example of content modification. We also compute the cosine similarity of textual webpage contents between Tor proxies and original ones using TF-IDF vectors. We provide the heatmap of text similarity for selected Tor proxies in each class of services in Fig. 10. We can observe that the similarity of the content of the webpages provided by many proxies is very low compared with the original ones, and some have different similarities under different services. This is probably because the proxies adopt different modifications for specific services (e.g., inclined to insert ads in forum services). In conclusion, we find that *almost all identified Tor proxies inject additional scripts/ads on the webpage of onion services or redirect to other pages*, which may raise potential risks for proxy users.

V. Discussion and Future Works

Ethics. We believe the measurements do not raise ethical concerns in this study. For the collection of PDNS records (for the measurement of §III), we only collect DNS response records in the PDNS records without any request record or

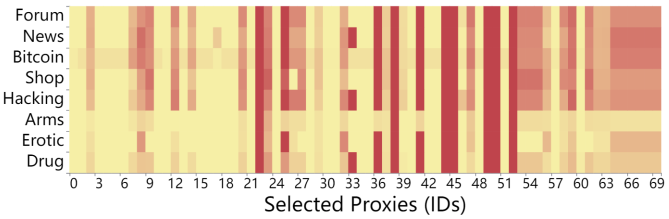


Fig. 10: The heatmap of text similarity (between proxy webpages and original ones) for selected Tor proxies in each class of services (*dark color means dissimilar*).

user information such as IP. Namely, only FQDN and resolved response are used. For the collection of webpages (for the measurement of §IV), we only crawl newly observed onion services each day, which is a total of thousands of times during two years. Such scanning is extremely slow and negligible for the crawled service. Meanwhile, we only collect textual content to avoid illegal materials of some onion services.

Our work is the first step of measuring Tor proxies throughout the Internet, which is initial but promising. We discuss limitations and several future directions below:

Finding More Proxies. In this work, we primarily identify Tor proxies via PDNS records. Compared with search engines that can only find a dozen of Tor proxies, we find ∼700 candidate Tor proxies. However, the number of proxies could still be improved in some directions: Firstly, more passive data and measurements can be involved. In this work, our data is collected from one of the most popular public DNS resolvers. Although it can widely cover the range of the Internet, data collected from a certain resolver may be geographically biased. Thus, future work could focus on introducing more passive data from more resolvers and investigating the impact of geographical bias on the measurement result. Secondly, more measurement methods can be involved, e.g., active measurement, as Tor proxies share similarities to HTTP proxies. How to efficiently scan and identify Tor proxies is a promising direction. Furthermore, another interesting question to investigate is how users learn about these proxies, as we can only find a few of them using search engines. This could further help us to find more proxies in a similar way. Moreover, our study focuses on analyzing Tor proxies based on the resolution of .onion domains via subdomains (like Tor2web). Other implementations that do not explicitly request the subdomains (e.g. passing the onion service's domain as an HTTP parameter) are out of scope in this study and can be explored in the future.

Security and Privacy Analysis. We conduct an initial measurement of content modification by Tor proxies in §IV-C. Future work can conduct case studies on the abnormal behaviors and other risks of Tor proxies. For example, whether and how certain proxies steal user privacy (e.g., bitcoin accounts and passwords as revealed in [37], [38]). Besides, since users are without an anonymity guarantee when using Tor proxies, censorship is available for governments. The organization behind Tor proxies and their actual purpose (whether malicious or censorship) also could be explored.

Use case Analysis. There are some potential use cases for Tor proxies beyond accessing onion services. For example, a use case is to track the emergence of new onion services. Future work could focus more on such analysis. Note that, such use cases depend on the large number of Tor proxies. The identified proxies and the identification method presented in our work could serve as the basis for further analysis.

VI. Conclusions

In this study, we conduct the first large-scale measurement study of Tor proxies throughout the Internet. Firstly, we cooperate with one of the most popular public DNS resolvers and collect passive DNS data with hundreds of billions of raw DNS records in four years, and propose techniques to identify and validate Tor proxies. We find ∼700 candidate Tor proxies in over 30 countries, of which 133 are valid (currently online after validation). Several analyses on the identified Tor proxies, including their lifespan, geographical distribution, domain names, and accessing scale are conducted and revealed. Secondly, we further analyze the usage of Tor proxy from the user side through the classification of onion services accessed via proxies and the proxy side through observing content modification of webpages from Tor proxies. We crawl newly observed webpages of onion services accessed via Tor proxies for two years. From the measurements, we report users' interest in using proxies to access different categories of services, and reveal a shocking finding of the content modification behaviors of proxies—over 97% Tor proxies have modified content of onion service webpages, including injecting ads or privacy-concerned scripts on the webpages or redirecting to other pages. We believe our first step of measurement could help to understand Tor proxies on the Internet and attract more future research attention to explore Tor proxies and Tor Network.

Acknowledgment

We are grateful to all the anonymous reviewers for their helpful comments. This work is supported by Tsinghua University (Institute for Network Sciences and Cyberspace) and QI-AIXIN Group Joint Research Center for Network Security, and also supported by Tsinghua University Initiative Scientific Research Program (2021Z11GHX010). Zhiliang Wang and Chenglong Li are the corresponding authors of this paper.

References

[1] P. Syverson, R. Dingledine, and N. Mathewson, "Tor: The secondgeneration onion router," in *Usenix Security*, pp. 303–320, 2004.

[2] T. T. P. Inc., "Tor metrics portal." https://metrics.torproject.org/. 2022.

[3] P. Winter, A. Edmundson, L. M. Roberts, A. Dutkowska-Żuk, M. Chetty, and N. Feamster, "How do tor users interact with onion services?," in *27th USENIX Security Symposium (USENIX Security 18)*, pp. 411–428, 2018.

[4] T. Project, "Tor2web: Browse the tor onion services." https://www.tor2web.org/. 2022.

[5] G. Pellerano, "Tor2web." https://github.com/tor2web/Tor2web.

[6] A. Mani, T. Wilson-Brown, R. Jansen, A. Johnson, and M. Sherr, "Understanding tor usage with privacy-preserving measurement," in *Proceedings of the Internet Measurement Conference 2018*, pp. 175–187, 2018.

[7] R. Jansen and A. Johnson, "On the accuracy of tor bandwidth estimation.," in *PAM*, pp. 481–498, 2021.

[8] T. Hoeller, M. Roland, and R. Mayrhofer, "On the state of v3 onion services," in *Proceedings of the ACM SIGCOMM 2021 Workshop on Free and Open Communications on the Internet*, pp. 50–56, 2021.

[9] R. Snader and N. Borisov, "A tune-up for tor: Improving security and performance in the tor network.," in *ndss*, vol. 8, p. 127, 2008.

[10] A. Greubel, S. Pohl, and S. Kounev, "Quantifying measurement quality and load distribution in tor," in *Annual Computer Security Applications Conference*, pp. 129–140, 2020.

[11] 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.

[12] N. Christin, "Traveling the silk road: A measurement analysis of a large anonymous online marketplace," in *Proceedings of the 22nd international conference on World Wide Web*, pp. 213–224, 2013.

[13] M. Bernaschi, A. Celestini, and G. Stefano, "Exploring and analyzing the tor hidden services graph," *ACM Transactions on the Web (TWEB)*, vol. 11, no. 4, pp. 1–26, 2017.

[14] S. Ghosh, A. Das, P. Porras, V. Yegneswaran, and A. Gehani, "Automated categorization of onion sites for analyzing the darkweb ecosystem," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1793–1802, 2017.

[15] 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*, pp. 35–43, 2017.

[16] 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)*, pp. 188–193, IEEE, 2014.

[17] G. Owen and N. Savage, "Empirical analysis of tor hidden services," *IET Information Security*, vol. 10, no. 3, pp. 113–118, 2016.

[18] D. Moore and T. Rid, "Cryptopolitik and the darknet," *Survival*, vol. 58, no. 1, pp. 7–38, 2016.

[19] N. S. Evans, R. Dingledine, and C. Grothoff, "A practical congestion attack on tor using long paths.," in *USENIX Security Symposium*, pp. 33–50, 2009.

[20] S. L. Blond, P. Manils, C. Abdelberi, M. A. D. Kaafar, C. Castelluccia, A. Legout, and W. Dabbous, "One bad apple spoils the bunch: exploiting p2p applications to trace and profile tor users," *arXiv preprint arXiv:1103.1518*, 2011.

[21] L. Overlier and P. Syverson, "Locating hidden servers," in *2006 IEEE Symposium on Security and Privacy (S&P'06)*, pp. 15–pp, IEEE, 2006.

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

[23] A. Kwon, M. AlSabah, D. Lazar, M. Dacier, and S. Devadas, "Circuit fingerprinting attacks: Passive deanonymization of tor hidden services," in *24th USENIX Security Symposium (USENIX Security 15)*, pp. 287–302, 2015.

[24] T. Wang and I. Goldberg, "Improved website fingerprinting on tor," in *Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society*, pp. 201–212, 2013.

[25] J. Hayes and G. Danezis, "k-fingerprinting: A robust scalable website fingerprinting technique," in *25th USENIX Security Symposium (USENIX Security 16)*, pp. 1187–1203, 2016.

[26] D. Arp, F. Yamaguchi, and K. Rieck, "Torben: A practical side-channel attack for deanonymizing tor communication," in *Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security*, pp. 597–602, 2015.

[27] 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*, pp. 1251–1260, 2017.

[28] D. Brown, "Resilient botnet command and control with tor," *DEF CON*, vol. 18, p. 105, 2010.

[29] O. Catakoglu, M. Balduzzi, and D. Balzarotti, "Attacks landscape in the dark side of the web," in *Proceedings of the Symposium on Applied Computing*, pp. 1739–1746, 2017.

[30] K. Li, P. Liu, Q. Tan, J. Shi, Y. Gao, and X. Wang, "Out-of-band discovery and evaluation for tor hidden services," in *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, pp. 2057–2062, 2016.

[31] M. Bernaschi, A. Celestini, S. Guarino, F. Lombardi, and E. Mastrostefano, "Spiders like onions: on the network of tor hidden services," in *The World Wide Web Conference*, pp. 105–115, 2019.

[32] J. Park, H. Mun, and Y. Lee, "Improving tor hidden service crawler performance," in *2018 IEEE Conference on Dependable and Secure Computing (DSC)*, pp. 1–8, IEEE, 2018.

[33] "114dns." https://www.114dns.com/. 2023.

[34] "Namesilo." https://www.namesilo.com/?rid=3853d48oc. 2022.

[35] J. Ramos *et al.*, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, vol. 242, pp. 29–48, Citeseer, 2003.

[36] G. Tsirantonakis, P. Ilia, S. Ioannidis, E. Athanasopoulos, and M. Polychronakis, "A large-scale analysis of content modification by open http proxies.," in *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2018.

[37] Caleb, "Tor2web proxies are using google analytics to secretly track users." https://medium.com/@c5/tor2web-proxies-are-using-google-analytics-to-secretly-track-users-fd245dbc81c5. 2018.

[38] M. Traudt, "Don't debug with onion.to." https://matt.traudt.xyz/posts/2021-12-02-dont-debug-with-onionto/. 2016.

Appendix A
Geographical Distribution of Tor Proxies

In Table VI, we provide the full list of geographical information of Tor proxies as the extension of Table III mentioned in §III-C, including the number of historically valid and online proxies.

Appendix B
Total Request Number of Tor proxies

We provide the total number in the last four months of PDNS records w.r.t. requesting onion services via each Tor proxy in Fig. 11. We can observe that the request scale of different proxies is imbalanced and long-tailed. This phenomenon is easy to understand and common on the Internet as some well-known proxies are widely used while many small proxies are not well promoted, which is also why it is difficult to find a comprehensive list of proxies as we mentioned before.

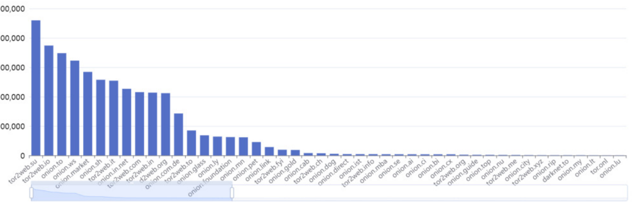
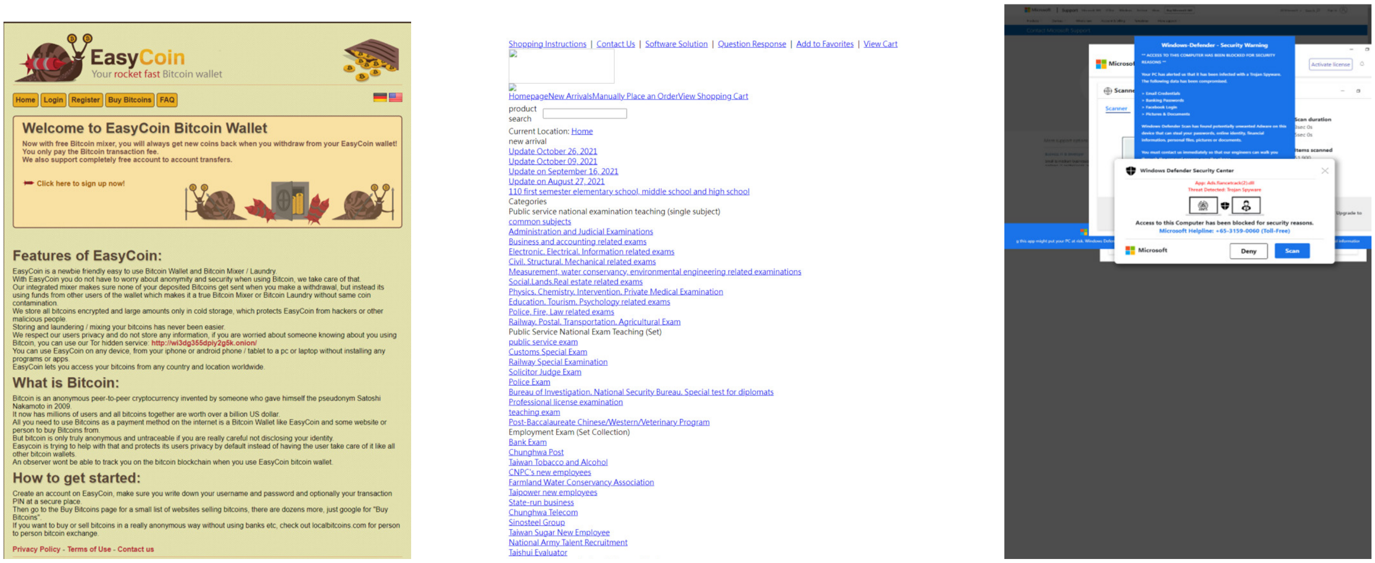


Fig. 11: The total number of requesting onion services via each Tor proxy.

Appendix C
Supplement of Service Classification.

Class Description and Keyword List. Here we provide supplement of service classification in §IV-B, including the class name and detailed description are in Table VII and selected keyword list for each class (for manual correction after machine learning clustering) is listed in Table VIII.

Classification Performance. We use a service list extracted from hidden wiki (as they have labels of service class) as a validation set for measuring the classification performance.



EasyCoin
Your rocket fast Bitcoin wallet

Home | Login | Register | Buy Bitcoins | FAQ

Welcome to EasyCoin Bitcoin Wallet

Now with free Bitcoin mixer, you will always get new coins back when you withdraw from your EasyCoin wallet!
You only pay the Bitcoin transaction fee.
We also support completely free account to account transfers.

Click here to sign up now!

Features of EasyCoin:

EasyCoin is a newbie friendly easy to use Bitcoin Wallet and Bitcoin Mixer / Laundry.
With EasyCoin you do not have to worry about anonymity and security when using Bitcoin, we take care of that.
Our integrated mixer makes sure none of your deposited Bitcoins get sent when you make a withdrawal, but instead its using funds from other users of the wallet which makes it a true Bitcoin Mixer or Bitcoin Laundry without same coin contamination.
We store all bitcoins encrypted and large amounts only in cold storage, which protects EasyCoin from hackers or other malicious people.
Storing and laundering / mixing your bitcoins has never been easier.
We respect our users privacy and do not store any information, if you are worried about someone knowing about you using Bitcoin, you can use our Tor hidden service: http://wi3dg355dply2g5k.onion/
You can use EasyCoin on any device, from your iphone or android phone / tablet to a pc or laptop without installing any programs or apps.
EasyCoin lets you access your bitcoins from any country and location worldwide.

What is Bitcoin:

Bitcoin is an anonymous peer-to-peer cryptocurrency invented by someone who gave himself the pseudonym Satoshi Nakamoto in 2009.
It now has millions of users and all bitcoins together are worth over a billion US dollar.
All you need to use Bitcoins as a payment method on the internet is a Bitcoin Wallet like EasyCoin and some website or person to buy Bitcoins from.
But bitcoin is only truly anonymous and untraceable if you are really careful not disclosing your identity.
Easycoin is trying to help with that and protects its users privacy by default instead of having the user take care of it like all other bitcoin wallets.
An observer wont be able to track you on the bitcoin blockchain when you use EasyCoin bitcoin wallet.

How to get started:

Create an account on EasyCoin, make sure you write down your username and password and optionally your transaction PIN at a secure place.
Then go to the Buy Bitcoins page for a small list of websites selling bitcoins, there are dozens more, just google for "Buy Bitcoins".
If you want to buy or sell bitcoins in a really anonymous way without using banks etc, check out localbitcoins.com for person to person bitcoin exchange.

Privacy Policy - Terms of Use - Contact us

(a) Original webpage.

Shopping Instructions | Contact Us | Software Solution | Question Response | Add to Favorites | View Cart
HomepageNew ArrivalsManually Place an OrderView Shopping Cart
product search
Current Location: Home
new arrival
Update October 26, 2021
Update October 09, 2021
Update on September 16, 2021
Update on August 27, 2021
110 first semester elementary school, middle school and high school
Categories
Public service national examination teaching (single subject)
common subjects
Administration and Judicial Examinations
Business and accounting related exams
Electronic, Electrical, Information related exams
Civil, Structural, Mechanical related exams
Measurement, water conservancy, environmental engineering related examinations
Social,Lands,Real estate related exams
Physics, Chemistry, Intervention, Private Medical Examination
Education, Tourism, Psychology related exams
Police, Fire, Law related exams
Railway, Postal, Transportation, Agricultural Exam
Public Service National Exam Teaching (Set)
public service exam
Customs Special Exam
Railway Special Examination
Solicitor Judge Exam
Police Exam
Bureau of Investigation, National Security Bureau, Special test for diplomats
Professional license examination
teaching exam
Post-Baccalaureate Chinese/Western/Veterinary Program
Employment Exam (Set Collection)
Bank Exam
Chunghwa Post
Taiwan Tobacco and Alcohol
CNPC's new employees
Farmland Water Conservancy Association
Taipower new employees
State-run business
Chunghwa Telecom
Sinosteel Group
Taiwan Sugar New Employee
National Army Talent Recruitment
Taishui Evaluator

(b) tor2web.xyz.to

Windows-Defender - Security Warning

** ACCESS TO THIS COMPUTER HAS BEEN BLOCKED FOR SECURITY REASONS **

Your PC has alerted us that it has been infected with a Trojan Spyware. The following data has been compromised.

> Email Credentials
> Banking Passwords
> Facebook Login
> Pictures & Documents

Windows Defender Scan has found potentially unwanted Adware on this device that can steal your passwords, online identity, financial information, personal files, pictures or documents.

You must contact us immediately so that our engineers can walk you

Windows Defender Security Center

App: Ads.fiancetrack(2).dll
Threat Detected: Trojan Spyware

Access to this Computer has been blocked for security reasons.
Microsoft Helpline: +65-3159-0060 (Toll-Free)

Microsoft

(c) hiddenservice.net

Fig. 12: Example of content modification by Tor proxies.

TABLE VI: The number of identified proxies by countries.

Country Name	# Candidate Proxies	# Valid Proxies
United States	200	82
China	98	14
Germany	30	12
British Virgin Islands	21	1
Singapore	14	2
France	12	7
South Africa	8	1
Netherlands	8	4
Australia	8	0
Nepal	7	0
Canada	7	0
Japan	5	0
Romania	5	0
Belize	5	0
India	4	1
Luxembourg	3	0
Ireland	3	1
South Korea	3	3
Moldova	2	0
Seychelles	2	1
United Arab Emirates	2	0
Switzerland	1	1
Denmark	1	0
Russia	1	0
United Kingdom	1	0
Ukraine	1	0
Finland	1	0
Iran	1	0
Kazakhstan	1	1
Lithuania	1	1
Egypt	1	0

With our classification method (machine learning and manual correction), only two services are wrongly classified among over 200 services (the accuracy is over 99%), which demonstrates the effectiveness of our classification method.

Content Modification. Here is an example of content modification mentioned in §IV-C. We show the original webpage of a Bitcoin onion service and the webpages modified by two Tor proxies in Fig. 12. We can see that Fig. 12b redirects to a shop store and Fig. 12b redirects to a phishing website.

TABLE VII: The name and description of each class of onion services.

Class	Detailed Description
Drug	Trading of drugs including kush, heroin, etc
Erotic	Provider of erotic service or porn vedio and graph.
Arms	Trading of firearms and weapons.
Hacking	Provider of hack service or new hack information
Shop	Trading of illegal things such as passport (except drug and arm)
News	Provider of news that is prohibited or censored on the Internet
Forum	Forum for user to discuss and personal blog
Bitcoin	Service about bitcoin including bitcoin wallet, mixer, etc
Inactive	Inactive service caused by unstable tor network or onion address update
Other	Clusters with small numbers of services and do not belong to the above classes

TABLE VIII: Selected keyword list for matching each class of onion services.

Class	Selected Keyword List*
Drug	[’drug’, ’he**in’, ’chr**ic’, ’k**h’, ’co**ne’, ’cry**al’, ’op**m’, ’mor**ia’,]
Erotic	[’porn’, ’sex’, ’erotic’, ’nude’, ’eja***ate’, ’r**e’, ’c*m’, ’blo**ob’, ’han**ob’, ’c**k’]
Arms	[’arm’, ’gun’, ’weapons’, ’P99’, ’P226’, ’PPK’, ’Glock’, ’ammo’, ’Bullet’]
Hacking	[’hack’, ’hacking’, ’hacker’, ’DDOS’, ’Exploits’, ’Phishing’, ’cracker’]
Shop	[’store’, ’market’, ’shop’, ’buy’, ’purchase’, ’sell’, ’sale’, ’credit’, ’retailers’, ’price’]
News	[’news’,]
Forum	[’forum’, ’board’, ’chan’, ’chat’, ’posts’, ’blog’]
Bitcoin	[’bitcoin’, ’blockchain’, ’chain’, ’mixer’]
inactive	[’error’, ’504’, ’404’, ’301’, ’502’]

[*] *We use asterisks to blur out some explicit words to avoid discomfort for readers.*