Computer Networks 208 (2022) 108893

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

Shining a light on dark places: A comprehensive analysis of open proxy ecosystem

Rui Bian [a,*], Shuai Hao [b], Haining Wang [c], Chase Cotton [a]

[a] Department of Electrical and Computer Engineering, University of Delaware, United States of America
[b] Department of Computer Science, Old Dominion University, United States of America
[c] Department of Electrical and Computer Engineering, Virginia Tech, United States of America

ARTICLE INFO

Keywords:
Open proxy
Content modification
Network measurements

ABSTRACT

Open proxies provide free relay services and are widely used to anonymously browse the Internet, avoid geographic restrictions, and circumvent censorship. To shed light on the ecosystem of open proxies and characterize the behaviors of open proxies, we conduct a large-scale, comprehensive study on over 436 thousand identified proxies, including 104 thousand responsive proxies in nine months. We characterize open proxies based on active and passive measurements and examine their network and geographic distributions, performance, and deployment. In particular, to obtain a more in-depth and broader understanding of open proxies, we analyze two particular groups of open proxies — cloud-based proxies and long-term proxies. To process and analyze the enormous amount of responses, we design a lightweight method that classifies and labels the proxies based on DOM structure which defines the logical structure of Web documents. We identify that 7.17% of responsive proxies modify the page content, and 76.42% of those proxies perform malicious actions. Furthermore, we parse the contents to extract information to identify the owners of proxies and track their activities for deploying malicious proxies. To this end, we reveal that some owners regularly change the proxy deployment to avoid being blocked and deploy more proxies to expand their malicious attacks.

1. Introduction

Open proxies provide free relay services to users, allowing them to browse the Internet anonymously [1,2], avoid geographic restrictions [3,4], or circumvent censorship [5–7]. Many open proxy aggregators [8–22] collect and publish thousands of "active" open proxies each day. Those enormous numbers of proxies have formed a large and complex ecosystem. In recent years, researchers have conducted studies to explore and characterize the open proxies in various aspects, such as performance, behaviors, security, and distributions [23–27]. They analyzed how the proxies can modify or manipulate the requested resources, such as HTML contents, image files, and executable files. The behaviors of such modifications have been used for advertisement injection [28–30], tracking user information [31,32], and malicious code execution [33,34]. However, the owners of those malicious proxies and corresponding campaigns have not been well studied before. In particular, open proxy owners can deploy and manage many proxies in diverse locations at different times to enhance the effectiveness of their activities or campaigns. Also, they could change their deployment and behaviors to hide their activities and avoid being detected and blocked. Thus, a systematic investigation on how open proxies are deployed and managed on the Internet is sorely needed but still missing.

In this paper, we perform a large-scale, comprehensive measurement-based analysis to investigate the ecosystem of open proxies. We design a measurement methodology to facilitate the analysis of massive returned responses from open proxies and accurately identify the proxies that manifest similar behaviors, possibly controlled by the same owner, to create a campaign. Moreover, to advance the understanding of the open proxy ecosystem, we study two specific groups of open proxies, the cloud-based proxies and long-term proxies. We identify and characterize the cloud-based open proxies by compiling a comprehensive list of cloud providers' IP ranges. We compare cloud-based open proxies with non-cloud-based open proxies in various ways. Open proxies are vulnerable to being abused due to their openness. As a result, typically the malicious open proxies could be quickly blacklisted [35] as their malicious behaviors are not hard to detect, and hence the lifetime of malicious open proxies is usually short. Therefore, to understand the usage and deployment of those long-term open proxies, we investigate the long-term proxies and compare them with short-term open proxies.

The three major contributions of this work are summarized as follows:

* Corresponding author.
E-mail addresses: bianrui@udel.edu (R. Bian), shao@odu.edu (S. Hao), hnw@vt.edu (H. Wang), ccotton@udel.edu (C. Cotton).

https://doi.org/10.1016/j.comnet.2022.108893
Received 13 September 2021; Received in revised form 15 January 2022; Accepted 5 March 2022
Available online 18 March 2022

- We collect more than 436 thousand open proxies in nine months, among which we identify and measure more than 104 thousand proxies that returned responses. To the best of our knowledge, the measurement scale of our work is the largest in the studies of open proxy in terms of data collection and analysis.
- We design a lightweight method to classify these open proxies based on the Document Object Model (DOM) structure. More importantly, we attempt to parse and extract the owner information of proxies that could be inferred from the HTTP responses. Through the analysis of malicious proxy owners, we discover different malicious cases and campaigns using open proxies. We further show that some owners are changing their deployments to avoid being blocked and deploy more proxies to enhance the power of their malicious attacks.
- We present an in-depth analysis of two specific deployments of open proxies, *i.e.*, the cloud-based open proxies and long-term open proxies. We study the characteristics of cloud-based proxies, showing that the cloud-based proxies have better performance and longer lifetime than non-cloud proxies. The cloud-based proxies also have a higher percentage of unchanged proxies for providing more reliable relay services. We also examine the long-term open proxies and uncover why they can survive in the wild Internet for a long time.

The remainder of this paper is organized as follows. We introduce the background of open proxy and survey the related work in Section 2. We present our methodology of measuring and analyzing open proxies in Section 3. We characterize the open proxy ecosystem in Section 4. In Section 5, we analyze the content modifications of open proxies and examine the owners and campaigns of malicious open proxies. Then, we study two special groups of open proxies, cloud-based proxies and long-term proxies, in Sections 6 and 7, respectively. In Section 8, we discuss the ethical considerations and limitations of this work. Finally, we conclude the paper in Section 9.

2. Background and related work

2.1. Background

A web proxy is a relay server that forwards HTTP(S) requests and returns responses between a client and a server. Generally, a web proxy allows a certain group of users to access web pages to reduce bandwidth or bypass geographic restrictions. In particular, open proxies are publicly available proxy servers that any user can use without authentication, simply configuring the corresponding IP address and port.

In many cases, open proxies can help users hide their original IP addresses to circumvent the geolocation-based restraint since the web-server can only see the open proxy's IP address. In contrast, some open proxies may reveal original IP addresses or the presence of the proxy by adding specific headers, such as `X-FORWARD-FOR` or `HTTP_VIA`.

2.2. Related work

Open proxy studies. Scott et al. [24] studied the open proxies that expose usage statistics from open management interfaces of manager programs such as Squid and analyzed the usage, distribution, and traffic pattern of identified open proxies. Tsirantonakis et al. [23] presented a study focusing on content modifications in open proxies by examining and comparing the DOM structure. They analyzed multiple types of malicious behavior, such as replacing advertisements, collecting user information, and fingerprinting browsers. Furthermore, Perino et al. [26] built an open proxy measurement platform to examine the characteristics, behavior, performance, and usage of open proxies. Mani et al. [25] also explored the availability, performance, HTML manipulation, and file manipulation of open proxies and compared open proxies with Tor. Choi et al. [27] conducted a comparative analysis of open proxies and residential proxies. They used passive methods to study open proxies' distributions, blacklist-check results and relations with GDP, Internet freedom, *etc.*In this study, we present a more comprehensive and larger-scale study of the open proxy ecosystem. More importantly, by identifying content modifications and malicious behavior, we attempt to extract the information that can be used to infer and track the open proxy owners who possibly control a bunch of proxies. Also, we first investigate two particular types of open proxies, cloud-based and long-term proxies.

Relay system studies. CoDeen [36,37] implemented a proxy network consisting of web cache servers deployed in PlanetLab and provided insights of the proxy system management and the analysis of unusual web traffic observed from the proxy view. Weaver et al. [38] proposed Netalyzr, a diagnostic tool to analyze the user's connections, and found that 14% of clients use a web proxy. Huang et al. [39] studied the presence of multiple types of middleboxes by leveraging the vantage points of residential IP proxy service. Mi et al. [40] explored the residential IP proxy ecosystem and its security and management issues.

Manipulations by middlebox. Chung et al. [41] detected end-to-end violations of DNS, HTTP, and HTTPS through a paid residential proxy service. They found that up to 4.8% of nodes are subject to some type of end-to-end violations. O'Neill et al. [42] measured the prevalence of TLS proxies using a probing tool deployed through Google AdWords campaigns. They found that 1 in 250 TLS connections are TLS-proxied and identified over 1000 malware interceptions. Carnavalet et al. [43] studied TLS proxies used by antivirus and parental control applications that would be vulnerable to Man-in-the-Middle attacks. Durumeric et al. [44] built a heuristic to detect HTTPS interception by characterizing the TLS Handshakes of popular browsers and interception products. Their study shows that TLS interceptions drastically reduce connection security. Tyson et al. [45] investigated HTTP header manipulation of proxies and middleboxes and analyzed the factors affecting head manipulation. In this study, we also examine and classify content modifications by open proxies.

3. Methodology

To have a broad view and deep understanding of the open proxy ecosystem, we systemically collect open proxies from multiple sources and test them using a website with static content under our control. We then detect content modification by DOM tree comparison of the original content and the proxied content. By combining information extraction with manual inspection, we classify modifications into different categories and identify malicious proxy owners who control a set of proxies that share the same behavior.

3.1. Collecting open proxies

In this study, we collect more than 436,000 open proxies in total from multiple sources, including:

- Websites that collect and publish open proxies,
- Open-source tools that collect, validate, and publish available open proxies,
- Crowd-sourcing open proxy lists published by users.

The details of collection sources are listed in Table 1. We collect open proxy information from the above sources daily in nine months (from September 2019 to June 2020). In particular, for several sources that update their lists hourly, we crawled them every hour. We compile proxies from all sources daily and remove duplicate proxies.

Table 1
Sources of open proxies.

Type of sources	Source
Proxy websites	proxy-daily [8] proxylistdaily [9] smallseotools [10] dailyfreeproxy [11] sinium [12] proxy-list.download [13] openproxy.space [14] proxyserverlist24 [15] live-socks [16]
Proxy collection tools	ProxyBroker [17] Gretronger Tool [46]
Other proxy lists	clarketm [18] TheSpeedX [19] opsxcq [20] fate0 [21] a2u [22]

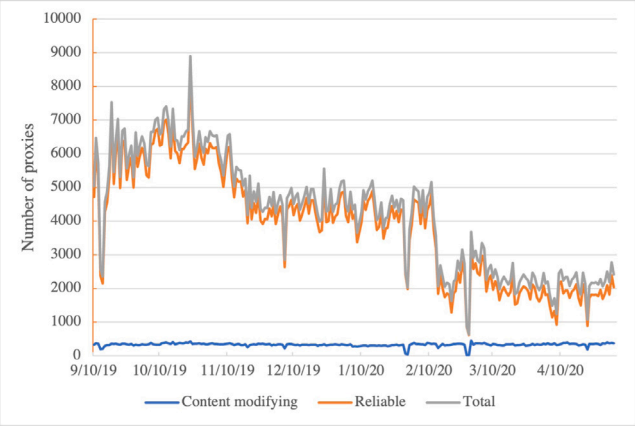


Fig. 1. Number of daily unique open proxies.

3.2. Measurement of open proxies

We conduct both active and passive measurements on collected open proxies to examine the open proxy ecosystem and behaviors.

Active measurement. To study the performance and behavior of open proxies, we set up two controlled websites and send HTTP/HTTPS requests to our controlled websites via each collected proxy. We simultaneously test 100 proxies and set 15 s timeout to filter out unresponsive or unreachable proxies. We use a server deployed in our university to issue requests to the static websites via proxies.[1] In each test, we record status code, response time (time from sending requests to receiving responses), download time (time from sending requests to finishing download all the requested resources), HTTP response headers, and HTTP page contents. In addition, to measure the performance, we send three ping probes from the deployed server to obtain the round-trip time (RTT) and use the `curl` to download a 5 MB test file via open proxies and to measure download speed.

Passive measurement. To understand the deployment and ecosystem of open proxies, we collect different types of data through diverse sources. Open proxies may have domain names associated with their IP addresses, so we perform the reverse DNS resolution (rDNS) to acquire domain names. To explore the distribution of open proxy networks, we query the WHOIS Database for AS information. Country-level geolocation of proxy is achieved by the Maxmind database [47]. To study the cloud proxies, we manually collect IP address ranges from 31 public cloud service providers to identify the proxies deployed in cloud platforms. Finally, we identify the blacklisted open proxies by leveraging the open-source blacklist scan tool `Pydnsbl` [48] that integrates data from 53 blacklist sources.

3.3. Detecting content modification and identifying open proxy owners

We employ a similar approach to detecting and clustering the content modification as the study done by Tsirantonakis et al. [23]. Specifically, we extract the DOM structure of returned content from proxies and compare it with the original web page's DOM structure. Although it is straightforward and convenient to detect modification or unexpected response by DOM structure, it is challenging to process massive data from thousands of proxies with modified contents. In total, we receive 83,815 unique response contents. We observe that proxy owners can change the modified contents by injecting or replacing them with random text in contents, but their DOM structures remain the same. To facilitate data processing, we first cluster content modification proxies to groups based on their DOM structures. Overall, we identify 1745 unique DOM structures from all collected responses. Through examination of several cases in each group, we classify the open proxies as benign or malicious. Furthermore, to identify possible owners of open proxy groups, we parse received HTML contents and extract elements, including metadata (title, keywords, and other fields), inject library, and URLs to search for identifiers of owners.

For the obfuscated codes, we manually inspect them by using multiple methods, including Unicode decoding, Base64 decoding, function evaluation, variable evaluation, and code formatting. By combining extracted elements and manual inspection, we can classify malicious behavior and identify open proxy group owners (detailed cases examined in Section 5.2).

4. Overview of open proxy characterization

In this section, we characterize the open proxy ecosystem. First, we present the network distribution and geographic distribution of open proxies. Next, we study the reliability and performance of *responsive* proxies. For content modifications and malicious owners, we present details in Section 5.

Daily statistics of proxies. The number of unique proxies (content modifying, reliable and total responsive proxies), over time, is shown in Fig. 1. The median number of daily reliable proxies is 4141.5, with a range of [622, 8473]. The median number of daily content modifying proxies is 337, with a range of [18, 452]. The responsive proxies include reliable and content modifying proxies. The median number of daily total responsive proxies is 4461.5, with a range of [640, 8899]. With our nine-month collections and testing, we collect 436,451 unique proxies and 104,114 responsive proxies (23.97% of collected proxies).

Port Distribution. The port distributions of collected and responsive proxies are shown in Table 2. Port 9999, 8080, 3128, 80, and 8118 are the most popular ports in open proxies. In collected proxies, there are 14,239 proxies (3.26%) found to use multiple different ports. In responsive proxies, there are 4677 proxies (4.49%) found to use multiple different ports. One proxy is found to use 403 different ports in total during the nine months. Those observations demonstrate that open proxy owners may often change the web proxy port. The reason might be that switching ports can protect the proxy server as malicious users cannot easily leverage the proxy servers for malicious purpose.

Domain names. There are 130,435 unique domain names of collected proxies and 32,409 unique domain names of responsive proxies. The

[1] In this work, we deployed one vantage point in our laboratory. Based on the previous study [25], the behavior of proxies does not significantly vary with the different locations of the vantage points. We also did not observe different behaviors when utilizing additional vantage points.

Table 2
Port distributions of collected and responsive proxies.

All proxies			Responsive proxies		
Port	#	%	Port	#	%
9999	96,802	22.18%	9999	34,599	33.23%
8080	74,072	16.97%	8080	27,348	26.27%
4145	47,543	10.89%	3128	7534	7.24%
3128	18,988	4.35%	80	5767	5.53%
1080	17,746	4.06%	8118	2050	1.97%
80	13,580	3.11%	53 281	1256	1.21%
38 801	10,514	2.41%	8888	1077	1.03%
9000	9303	2.13%	8213	1025	0.98%
8118	8817	2.02%	3129	907	0.87%
8888	4158	0.95%	999	809	0.78%
All others	134,928	30.91%	All others	21,742	20.88%

Table 3
Domain name distributions of collected/responsive proxies.

Domain name	Count	Percentage
All proxies		
NXDOMAIN	229,481	63.07%
hn.kd.ny.adsl.	1078	0.3%
azteca-comunicaciones.com.	325	0.09%
static.vnpt.vn.	220	0.06%
int0.client.access.fanaptelecom.net.	164	0.05%
All others	132,255	36.43%
Responsive proxies		
NXDOMAIN	60,906	64.57%
azteca-comunicaciones.com.	177	0.19%
hn.kd.ny.adsl.	111	0.12%
static.vnpt.vn.	82	0.09%
customer.worldstream.nl.	52	0.06%
All others	32,859	34.97%

most common domain names resolved from the IP addresses of collected and responsive proxies are shown in Table 3. More than 60% of reverse DNS lookup results is `NXDOMAIN`, which means those proxies do not have domain names. Because users only need the IP address and port to use open proxies, it is reasonable that open proxies do not possess domain names necessarily. In addition, we manually inspect other popular names associated with open proxies and find that many of them have been noticed by their abnormal behaviors:

- `hn.kd.ny.adsl` often changed its matching IP address and those IP address belong to China Unicom. This domain name is reported to perform repetitive port scans and blind SQL injections [49–54]. In addition, because we use reverse DNS lookup to find the domain name of open proxies, the returned results might not be the real domain names of open proxies. `hn.kd.ny.adsl` is not a valid fully qualified domain name (FQDN), and we speculate that it is an internal domain name leaked to the public.
- `azteca-comunicaciones.com` is the domain name of a Columbia communication company — Azteca Comunicaciones. It also has been found to be mapped to many IP addresses and those IP addresses are identified as open proxies and spammers [55–57].
- `static.vnpt.vn` matches multiple IP addresses and all of them belong to VietNam Data Communication Company. This domain name is reported to send spams through different IP addresses [58–61].

Geolocation. The geolocation information of collected open proxies is shown in Table 4 and Fig. 2. The collected proxies are located in 172 countries, and the geographic distributions are skewed that over 80% of open proxies are located in 10 countries. China, Thailand, United States, Brazil, India, and Indonesia have the most collected open proxies and responsive proxies.

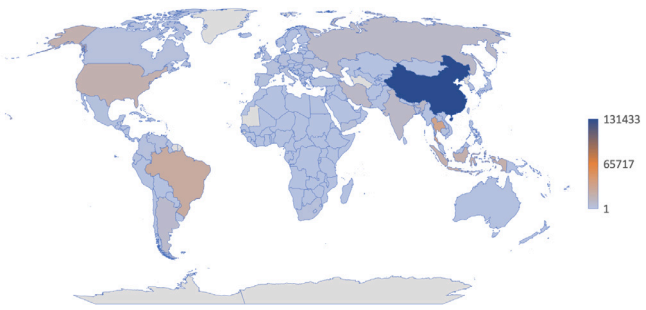


Fig. 2. Geo-distribution of open proxies.

Table 4
Geolocation of collected and responsive proxies.

All proxies		Responsive proxies	
Country	%	Country	%
China	41.92%	China	38.15%
Thailand	8.70%	Thailand	8.56%
United States	7.32%	Indonesia	7.89%
Brazil	6.14%	United States	6.90%
Indonesia	5.76%	India	5.04%
India	3.21%	Brazil	4.88%
Iran	3.03%	Russia	3.20%
Russia	2.77%	Iran	1.36%
Argentina	2.02%	Singapore	1.15%
Ukraine	1.20%	Bangladesh	1.14%
All others	17.92%	All others	21.69%

Table 5
Content modifications of proxies.

Behavior	# Proxy	Percentage
Always modify	6326	6.04%
Never modify	97,074	92.73%
Sometimes modify	1287	1.23%

Cloud. In collected proxies, 18,005 proxies (4.13%) are hosted on the public cloud platform. In responsive proxies, 5637 proxies (5.41%) are hosted on public cloud platforms. The details of the cloud-based open proxy study are presented in Section 6.

Autonomous System (AS). The collected proxies reside in 9060 ASes, and responsive proxies reside in 5282 ASes. The most popular ASes for collected and responsive proxies are shown in Table 7. The distributions of AS are also significantly unbalanced, where more than half of open proxies reside in only ten ASes. Most of these ASes belong to telecommunication and Internet companies that provide server hosting services.

Blacklist. The open-source blacklist scan tool `Pydnsbl` [48] that integrates data from 53 sources is used to extract open proxies being blacklisted. In collected proxies, 272,719 proxies (62.48%) appear in at least one blacklist, 163,732 proxies are not on any blacklist. In responsive proxies, 70,122 proxies (67.35%) appear in at least one blacklist, 33,992 proxies are not found on blacklists. The high percentage shows that most open proxies may have performed suspicious or malicious activities.

Behavior. The content modification results are shown in Table 5. We identify that 92.73% of proxies always returned the expected response all the time. This result shows that most of the working open proxies are reliable. In the meantime, 6.04% of proxies consistently perform the content modification. Interestingly, 1.23% of proxies change their behaviors from time to time. The owners of these proxies may change their behavior by purpose to hide their malicious activities and avoid

Table 6
Lifetime and performance of proxies.

Average	Responsive	Reliable	Modifying
Lifetime (days)	9.45	9.37	10.89
Response time (s)	4.99	5.24	1.95
Download time (s)	5.12	5.37	2.04
RTT (ms)	233.24	231.7	250.78
Download speed (KBps)	254.43	271.07	57.47

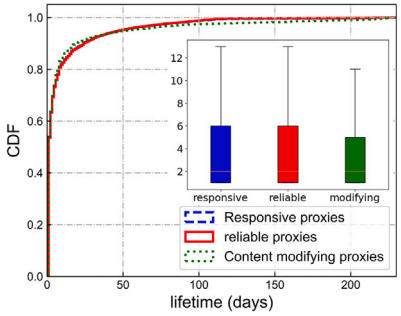


Fig. 3. CDF and boxplot of lifetime.

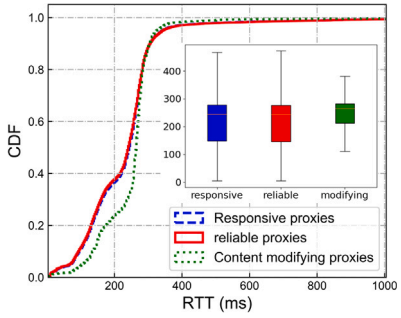


Fig. 4. CDF and boxplot of RTT.

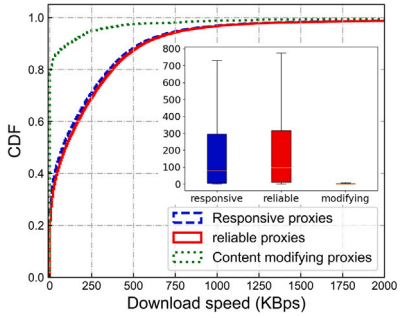


Fig. 5. CDF and boxplot of download speed.

being detected. We describe a detailed analysis of content modifications in Section 5.

Lifetime. Here, we further define the open proxies which consistently return unchanged content as reliable proxies. The CDFs and boxplots of proxy lifetimes (responsive, reliable, and content modification proxies) are shown in Fig. 3. The average lifetime of responsive, reliable, and content modification proxies are shown in Table 6. We observed that nearly 80% of proxies' lifetime is one week or less. Content modification proxies' lifetime is slightly longer than reliable proxies, which means content modification proxies are more resistant than reliable proxies. The detailed discussions about content modification proxies are presented in Section 5, and we discuss long-term proxy in Section 7.

Performance. The CDFs and boxplots and of responsive, reliable, and content modification proxies' RTTs and download speed are shown in Figs. 4 and 5. The performance of responsive, reliable, and content modification proxies is shown in Table 6. The figures and table show that reliable proxies have better performance than content modification proxies, with shorter RTTs and faster download speed.

Summary. In this section, we characterize open proxies from multiple aspects. We present network distributions (port, domain name, and AS), geographic distribution, lifetime, performance (RTT and download speed), and reliability (blacklist check and content modification). Moreover, we observed that the majority of open proxies are concentrated in a small set of AS and countries. The lifetime of open proxies is very short that most proxies cannot live up to one week. Two-thirds of open proxies are listed in blacklists, and 7.31% of open proxies returned modified contents.

5. Content modification and malicious open proxy owners

In this section, we identify the behaviors of malicious and benign proxies and present detailed case studies to explore the owners who deploy the malicious proxies and how the proxy owners can benefit from the campaigns using open proxies.

5.1. Content modification

Since we received thousands of responses via proxies every day, it is challenging to process and analyze such massive data. To reduce manual effort and simplify the analysis, we utilize the DOM structure to analyze the contents. To do so, we simply record the tag names and locations of each HTML contents. If there are different tag names or locations between two HTML pages, we consider those two DOM structures are different. In total, we identified 1745 unique DOM structures of all collected responses. Next, we select representative cases to classify proxies. We parse the HTML contents to extract proxy activity information to understand each proxy group's behavior and nature.

By combining extracted information with the manual examination, we classify the content modification proxies as benign or malicious. We consider the following scenarios with content modification proxies as benign:

- Lack of permission: access is denied due to no proper permission;
- Errors: that category includes network errors like DNS errors and configuration errors;
- Misclassification: incorrectly labeled as open proxies by open proxy collecting source;
- Blocked by network management software or AntiBot software, probably due to a restricted access policy.

Then, we identify the following cases of content modifications as the misbehavior of malicious proxies:

- Replacing original content: such proxies replace the static content in our original server and lead the user to other websites (shopping, adult, and news website) or applications;
- Ad injection: this type of proxies inject advertisement JavaScript to the original contents;
- CSS injection: these proxies inject the suspicious CSS file;
- Redirection: these proxies redirect users to other websites;
- Collecting user information: these proxies inject scripts to obtain user information like Operating System, browser, and cookie;
- Cryptojacking: these proxies inject cryptocurrency mining scripts that take advantage of the user's resource to mine digital currency by stealth.

Table 7
Most popular ASes for collected and responsive proxies.

All proxies			Responsive proxies		
ASN	Organizations	Percentage	ASN	Organizations	Percentage
4134	No. 31,Jin-rong Street	24.89%	4134	No. 31,Jin-rong Street	27.9%
37 963	Alibaba Advertising Co.,Ltd.	8.87%	4837	China Unicom China169 Backbone	5.87%
4837	China Unicom China169 Backbone	5.61%	14 061	DigitalOcean, LLC	3.43%
23 969	TOT Public Company Limited	3.37%	45 758	Triple T Internet/Triple T Broadband	3.36%
7713	PT Telekomunikasi Indonesia	3.01%	7713	PT Telekomunikasi Indonesia	3.03%
14 061	DigitalOcean, LLC	2.43%	23 969	TOT Public Company Limited	2.22%
45 758	Triple T Internet/Triple T Broadband	2.24%	17 816	China Unicom IP network China169	2.09%
131 090	CAT TELECOM Public Company Ltd	1.38%	17 552	True Internet Co.,Ltd.	1.24%
16 276	OVH	1.17%	20 473	Choopa, LLC	1.12%
17 552	True Internet Co.,Ltd.	1.06%	17 451	Biznet Networks	1.12%
	All others	45.99%		All others	48.61%

Table 8
Categories of content modification proxies.

	Category	# Proxy	Percentage
Benign (23.58%)	Lack of permission	1234	16.52%
	Error	112	1.50%
	Misclassification	366	4.90%
	Blocked	49	0.66%
Malicious (76.42%)	Replacement	466	6.24%
	Ad injection	2393	32.04%
	CSS injection	9	0.12%
	Redirection	2748	36.80%
	Collect user information	96	1.23%
	Cryptojacking	19	0.25%

The categories of benign and malicious proxies are shown in Table 8. We identified 23.58% of content modifications are benign, and the majority of them are due to lack of permission or misclassification. The possible reason is that open proxy collectors did not validate the nature and availability of collected proxies and public them incorrectly in open proxy lists to the Internet. Malicious proxies occupy 76.42% of content modification proxies. Most of the malicious proxies belong to two categories — Ad injection and redirection. In addition, we find 19 proxies performing cryptojacking attacks.

5.2. Malicious open proxy owners: Case studies

Open proxies offer service for users free of charge, but the deployment is not free for owners. To understand the purpose and benefit of deploying open proxies, we attempt to identify and track the open proxy owners by information parsed from modified contents as we find that some proxy owners typically deploy and control a set of proxies that perform the same modifications. In this part, we discuss several case studies to demonstrate the purposes and deployment of open proxies by means of their owners.

ISP injection. Many open proxies inject similar JavaScript code snippet to display advertisements or collect user's information for censorship. They obtain user's information including domain name, screen width and height, and other parameters like id, enc, params, and idc_r. These proxies label users by allocating different parameters like id and enc. These pieces of information are concatenated to two common URLs (`'notifa.info'` and `'cfs.uzone.id'`) and then sent back. The example of injected code by ISP is shown in Fig. 6.

In total, we identified 6572 such responses from 2107 proxies observed in 237 days. The most proxies observed in one day are 86 proxies, and the average proxies observed in one day is 27.73. The lifetime range of those proxies is from 1 day to 102 days. This group of proxies belong to 43 ASes, all located in Indonesia. Indonesia ISP hosts these proxies to sell the ads and censor the traffics. Even though it is not clear if they are illegal, it is better to avoid using these proxies to protect users' privacy.

```
<script type = "text/javascript" >
  if (self == top) {
    function netbro_cache_analytics(fn, callback) {
      setTimeout(function() {
        fn();
        callback();
      }, 0);
    }

    function sync(fn) {
      fn();
    }

    function requestCfs() {
      var idc_glo_url = (location.protocol == "https:" ?
"https://" : "http://");
      var idc_glo_r = Math.floor(Math.random() *
99999999999);
      var url = idc_glo_url + "p03.notifa.info/3fsmd3/request" +
"?id=1" + "&enc=9Uw…gY9" + "&params=" + "4Tt……%3d" +
"&idc_r=" + idc_glo_r + "&domain=" + document.domain +
"&sw=" + screen.width + "&sh=" + screen.height;
      var bsa = document.createElement('script');
      bsa.type = 'text/javascript';
      bsa.async = true;
      bsa.src = url;
      (document.getElementsByTagName('head')[0] ||
document.getElementsByTagName('body')[0]).appendChild(bsa);
    }
    netbro_cache_analytics(requestCfs, function() {});
  }; < /script>
```

Fig. 6. Injected code by ISP.

Cloud provider advertisement. We identified a group of proxies that inject JavaScript codes to provide Chinese cloud provider advertisements, called `Ruijieyun`, a cloud platform for marketing and third-party payment. The scripts will detect the user's IP address and determine if the user's IP is in their IP address ranges. If so, they will not provide ads, while if not, they will pop up ads to promote their cloud service. That strategy can enhance the ads' effectiveness to make ads only propagate among new users. In total, we received 4067 responses from 420 proxies observed in 221 days. The maximum number of such proxies observed in one day is eighty. The average of such proxies observed in one day is 18.40, with the lifetime ranging from 1 day to 73 days. They reside in 14 Chinese ASes that all belong to Chinese telecommunication Companies. This case shows that `Ruijieyun` cloud service

company deploys open proxies in multiple China telecommunication company networks to broadcast its advertisements for attracting new users to utilize its cloud service.

User network information collection. We observed that a group of proxies inject similar JavaScript code in the headers. These injections do not change the original contents but prompt users to send requests to the Google Analytics website with specific parameters:

```
https://www.google-analytics.com/collect?v=1&
t=pageview&tid=UAXXXXXXXXXXXX&dh=test777.com&
cid=XXXXX&dp=/mp/ping/
```

The actual user ID is marked here to protect privacy. We parse the URL and parameters based on the references of Google Analytics. We focus on four key parameters: `tid`, `dh`, `cid`, and `dp`. The `tid` is tracking ID or web property ID that is associated with collected data. The `dh` is document hostname that specifies the hostname from which content was hosted. The `cid` is client ID that is used to identify a particular user, device, or browser instance. The `dp` is document path which is the path portion of the page URL. In the collected data, `tid`, `dh` and `dp` are identical in this open proxy group, while the `cid` is changed in each request. All the cases share the same tracking ID, which indicates that all collected data associates with the same owner. In addition, this owner collected user information that is hosted in one particular hostname (`test777.com`) and document path. We visited this website to explore the owners' purpose and found a Japanese research website for network and hardware experimentation. It has stopped updating since 2006. We notice the document path is named as 'ping', which could imply PING probing measurements. We speculate that this owner collected network measurement data from users by injecting JavaScript code. In total, we observed 338 responses from 54 proxies. Those proxies are located in 11 countries in Europe, Asia, and North America, indicating that the owner has deployed a large number of widely distributed proxies to obtain a large amount of measurement data. However, we argue that the owners should well inform users of the measurement content and obtain users' consent to conduct measurements in such a large-scale experiment. Also, the experiment code should be cleared up if proxy owners discontinue the measurement.

Cryptojacking. We identified a group of open proxies performing cryptojacking. The contents they returned look like a regular login page of an online forum that requests a username and password. Meanwhile, they inject JavaScript code that pops out a YouTube video while using the user's processor to mine cryptocurrency without permission or notification. The screenshots of the cryptojacking page are presented in Fig. 7. By carefully inspecting the injected codes, we find that all the mining scripts contain the same identifier (*i.e.*, a wallet ID), which means all the mined cryptocurrencies will benefit the same owner. Hence, we infer that this owner deploys or rents multiple malicious proxies to enhance his/her mining capacity and obtain profit. In total, we received 1416 responses from 19 proxies observed in 106 days. The maximum number of such proxies observed in one day is nineteen. The average proxies observed in one day is 12.91. The most extended lifetime of them is 94 days, and the shortest lifetime is 38 days. The owners choose to use different proxies and change the number of proxies to avoid being detected and blocked. Also, 18 out of 19 observed proxies are hosted in AS 14061 Digital Ocean, a popular global cloud infrastructure provider, while one is in Hetzner — a German Internet hosting company. These proxies are distributed in seven countries in North America, Asia, and Europe. This type of malicious proxies could cause considerable damage to users because if users do not notice this video and leave this video open, these malicious proxies can take advantage of users' processors to mine cryptocurrency for a long time.

Ad injection campaigns. One group of proxies returned a mobile news application called Orange News — a Hong Kong news application. Some proxies will provide business websites such as `Early Bird`

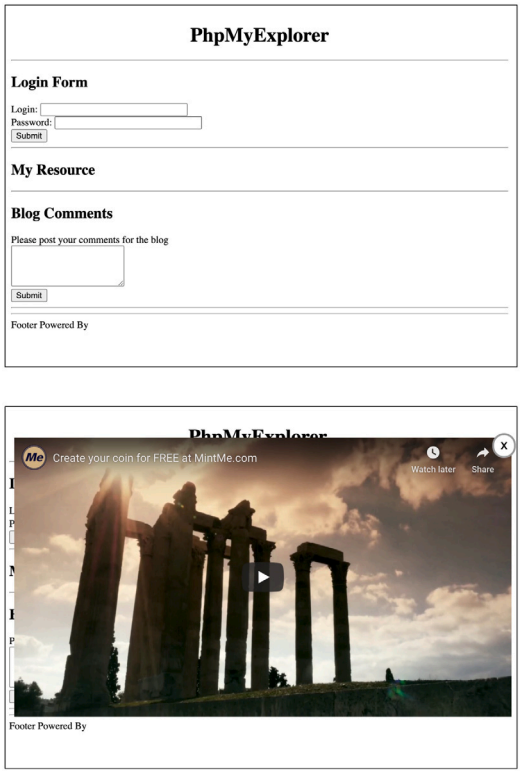


Fig. 7. Screenshots of the cryptojacking page. The top is the login page, and the bottom is the login page covered by the pop-up video of Ad.

`Cashflow`, which provides cash flow service, and `DragonEX` which offers digital currency trade and exchange service. In another case, proxies return a web game called `Tank Rumble` that users can use mouse and keyboards to control the tank to attack enemies. Another returned questionable content is an education website that provides an English training program called `Cambridge English`. Another application the proxies returned is a game communication app called `Nadeko`. No matter whether these websites own those proxies, it is reasonable to infer that those proxies' owners can obtain profit by redirecting users to their desired websites.

In this section, we first categorize open proxies based on the content modification behaviors. About 23.58% of open proxies that modify contents are benign, and most of them fall into the categories of the lack of permission and mis-classification. Then, we focus on the malicious open proxies that occupy 76.42% of content modification proxies. We conduct detailed case studies to thoroughly analyze the malicious open proxies' behaviors and deployments to explore proxy owners' purposes. The owners may achieve monetization from proxy users by injecting advertisements, collecting user information, replacing original content with applications and websites, and mining cryptocurrency. These proxy owners use open proxies to expand their influences and gain profits from numerous users.

6. Cloud-based open proxy

Cloud service has quickly grown in recent years, with 84% of organizations now using cloud services, up from a mere 48% five years earlier. In this section, we study open proxies hosted in the cloud, and we refer to them as cloud-based proxies. To identify the proxies hosted in cloud platforms, we first collect popular cloud providers' public IP address ranges from their official websites. In total, 31 cloud providers' public IP address ranges are collected. For large cloud providers like Amazon, Microsoft, and Google, we also collected the IP ranges of their

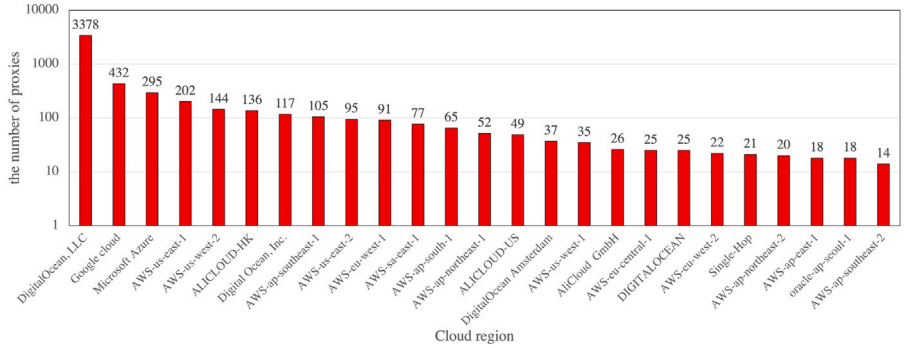


Fig. 8. The number of collected proxies in different regions of cloud platforms.

Table 9
Top 10 countries hosting cloud-based and non-cloud-based proxies.

Cloud-based proxy		Non-cloud-based proxy	
Country	Percentage	Country	Percentage
United States	59.89%	China	40.34%
Singapore	13.22%	Thailand	9.05%
Canada	5.57%	Indonesia	8.35%
United Kingdom	3.73%	India	5.20%
Netherlands	3.39%	Brazil	5.07%
Germany	2.77%	United States	3.87%
India	2.25%	Russia	3.39%
Ireland	1.84%	Iran	1.44%
Brazil	1.67%	Bangladesh	1.21%
Hong Kong	1.51%	Argentina	1.16%

regions. In total, we found 1733 cloud regions and 29,632 cloud IP address blocks. Then we verify whether responsive proxy IP addresses are in the cloud IP blocks and collect cloud-based proxies. There are 5637 responsive proxies in 57 cloud regions. The top 25 Cloud regions that contain proxies are shown in Fig. 8. Most cloud-based proxies belong to Digital Ocean, Google Cloud, Azure, and Amazon. Nearly 90% of cloud-based proxies belong to the top 10 cloud regions.

Geolocation. We present the top 10 countries of cloud-based proxy and non-cloud-based proxy in Table 9. Most cloud-based proxies are located in developed countries such as the US, Singapore, and the UK. Most of them are in North America and Western Europe. Besides, most non-cloud-based proxies are located in developing areas such as Asia and South America. That is perhaps because the Cloud services are more prevalent and available in developed countries than developing countries, and open proxy owners can quickly and inexpensively deploy their open proxy servers on the cloud. For developing countries, the cloud service is not widely available, and the price is relatively high, so open proxy owners unlikely to use the cloud services to deploy proxies. Also, more than 40% of non-cloud-based proxies are in China, and the main reason is likely that Chinese users may utilize open proxies to circumvent censorship.

Blacklist. The results of blacklist check for cloud-based proxy and non-cloud-based proxy are shown in Table 10. The percentage of proxies found in the blacklist is quite different: 69.39% of non-cloud-based proxies are found in the blacklists, while only 31.81% of cloud-based proxies blacklisted. The possible reasons for fewer proxies found in blacklist in the cloud are (1) cloud-based proxies are managed and monitored by cloud service providers, so they will be detected and blocked if they violate cloud service's policies; (2) the cloud-based proxies could be more dynamic than non-cloud-based proxies due to the elastic resource provision of cloud services, and blacklists are limited to detect such dynamic cloud IP addresses.

Behavior. Content modifications of cloud-based proxy and non-cloud-based proxy are shown in Table 11. The percentage of proxies that

Table 10
Blacklist check results of cloud-based and non-cloud-based proxies.

	Cloud-based proxy		Non-cloud-based proxy	
	#proxy	Percentage	#proxy	Percentage
In BL	1793	31.81%	68,329	69.39%
Not in BL	3844	68.19%	30,148	30.61%

Table 11
Content modifications by cloud-based and non-cloud-based proxies.

	Cloud-based proxy		Non-cloud proxy	
	#proxy	Perc.	#proxy	Perc.
Always modify	163	2.89%	6023	6.12%
Never modify	5393	95.67%	91,248	92.66%
Sometimes modify	81	1.44%	1206	1.22%

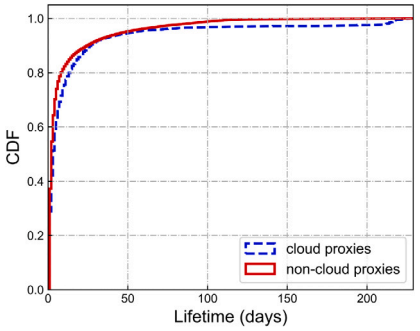


Fig. 9. CDF of lifetime of cloud-based proxy and non-cloud-based proxy.

constantly modify the contents of the cloud-based proxy (2.89%) is lower than that of non-cloud-based proxy (6.12%). Interestingly, the percentage of proxies that intermittently modify the contents of the cloud-based proxy (1.44%) is slightly higher than that of non-cloud-based proxy (1.22%). Due to the cloud's dynamic and elasticity, open proxy owners can easily manage and change the proxy settings and configurations, so they may adjust their policies to modify or just forward the contents. By combining the blacklist and behavior results, we can see that cloud-based proxies have better reliability than non-cloud-based proxies.

Lifetime. The CDF of cloud-based proxies and non-cloud-based proxies' lifetime is shown in Fig. 9. The average lifetime of cloud-based proxies and non-cloud-based proxies are shown in Table 12. We can see that most cloud-based proxies have a longer lifetime (14.19 days) than non-cloud-based proxies (9.17 days). Cloud infrastructures can provide more protection so that cloud-based proxies are more resistant than non-cloud-based proxies.

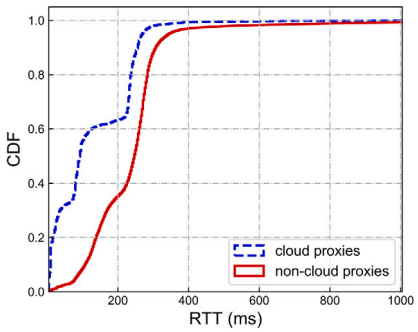


Fig. 10. CDF of RTT of cloud-based proxy and non-cloud-based proxy.

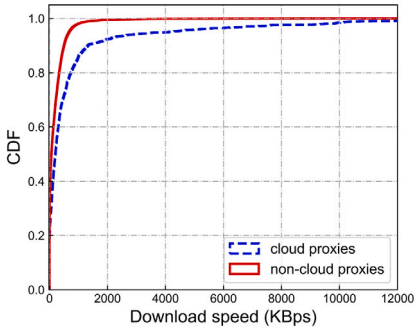


Fig. 11. CDF of download speed of cloud-based proxy and non-cloud-based proxy.

Table 12
Lifetime and performance of cloud-based and non-cloud-based proxies.

Average	Cloud-based	Non-cloud
Lifetime (days)	14.19	9.17
Response time (s)	4.28	5.04
Download time (s)	4.31	5.18
RTT (ms)	129.3	238.83
Download speed (KBps)	811.93	195.65

Performance. The CDF of cloud-based proxies and non-cloud-based proxies' RTT and download speed are presented in Figs. 10 and 11. The performance of cloud-based proxies and non-cloud-based proxies are shown in Table 12. These comparisons show that cloud-based proxies have better performance than non-cloud-based proxies. Cloud-based proxies have shorter RTT – near half of the non-cloud-based proxies' RTT, and cloud-based proxies have faster download speed – more than four times of non-cloud-based proxies' download speed. Typically, cloud service can provide better performance than traditional servers, so cloud-based proxies have better performance due to this reason.

In this section, we study a specific type of open proxies — the cloud-based proxies. We present cloud-based proxies' network and geographic distribution, behavior, and performance, and compare them with non-cloud-based proxies. We analyze the reasons causing the differences between cloud and non-cloud-based proxies. Even though the scale of cloud-based proxies is smaller than that of non-cloud-based proxies, cloud-based proxies have multiple advantages such as higher reliability and better performance over non-cloud-based proxies. Also, proxy owners can take advantage of the cloud to change the proxy's behavior and make cloud-based proxies more dynamic.

7. Long-term open proxy

It is easy and convenient to use open proxy since the proxy setting is simple (only enter the IP address and port) without authentication and free of charge. On the other hand, open proxies make it easier for miscreants to launch a variety of attacks. Hence, open proxies are vulnerable to be attacked and abused. To this end, open proxies' lifetime is relatively short. In our study, the average lifetime is 9.45 days. 53.93% of responsive proxies' lifetime is two days or less, and 80.92% of responsive proxies' lifetime is short than ten days. Only 0.20% responsive proxies' lifetime is more than two hundred days. Here, we examine the *long-term* proxies whose lifetime are equal and longer than two hundred days and compare them to relatively *short-term* proxies whose lifetime is less than ten days. In this section, we examine the characteristics of long-term open proxies and explore how and why they exist for quite a long time.

Table 13
Top 5 ASes of long-term and short-term proxies.

ASN	AS	Percentage
	Long-term proxies	
14 061	Digital Ocean	61.14%
39 832	Opera Software	5.69%
24 940	Hetzner Online GmbH	5.21%
16 509	Amazon.com, Inc	2.37%
37 963	Hangzhou Alibaba Advertising Co.,Ltd.	1.90%
	Short-term proxies	
4134	No. 31,Jin-rong Street	34.14%
4837	China Unicom China169 Backbone	7.10%
45 758	Triple T Internet/Triple T Broadband	4.04%
7713	PT Telekomunikasi Indonesia	3.65%
14 061	DigitalOcean, LLC	3.00%

Table 14
Top 10 countries of long-term proxy and short-term proxy.

Long-term proxy		Short-term proxy	
Country	Percentage	Country	Percentage
United States	36.49%	China	45.74%
Germany	9.95%	Thailand	9.53%
Netherlands	9.95%	Indonesia	7.21%
India	8.06%	United States	6.63%
Singapore	7.11%	India	4.24%
Canada	6.16%	Brazil	4.21%
United Kingdom	5.21%	Russia	2.32%
China	4.74%	Iran	1.35%
Russia	2.84%	Singapore	1.18%
Iran	1.42%	France	0.85%

Autonomous System. The top 5 ASes of the long-term and short-term proxies are shown in Table 13. Most long-term proxies are hosted in Digital Ocean. In contrast, most short-term proxies are hosted in telecommunication networks like China Telecom, China Unicom, Thailand Triple *T* Internet, and Indonesia PT Telkom. As the discussion in Section 6, cloud services provide elastic resources and more protections so that it is reasonable that most long-term proxies are deployed in ASes belonging to cloud platforms. We identify that 64.45% of long-term proxies are host in the cloud, while only 4.91% of short-term proxies are hosted in the cloud. The possible reason is that cloud service provides more reliable and resistant service to host proxy servers so that long-term proxies contain a higher percentage of cloud proxies.

Geolocation. We present the top 10 countries of long-term proxy and short-term proxy in Table 14. Most long-term proxies are distributed in developed countries such as the US, Germany, and the Netherlands. Most of them are in North America and Western Europe. Besides, most short-term proxies locate in less developing countries like China, Thailand, and Indonesia. The reason might be that open proxies in developing countries are more vulnerable due to strict control, including filtering and censorship.

Table 15
Content modifications of long-term and short-term proxy.

	Long-term	Short-term
Always modify	32.70%	6.20%
Never modify	67.30%	92.63%
Sometimes modify	0.00%	1.17%

Table 16
Performance of long-term and short-term proxy.

Average	Long-term	Short-term
Response time (s)	0.85	4.68
Download time (s)	0.86	4.82
RTT (ms)	119.56	238.06
Download speed (KBps)	901.56	238.04

Blacklist. The percentage of proxies found in blacklists is quite different: 69.05% of short-term proxies are blacklisted, while only 18.01% of long-term proxies are included by those blacklists.

Behavior. Content modifications of long-term proxy and short-term proxy are presented in Table 15. Even though the modification rate of long-term proxy is higher than short-term proxies, after analyzing the categorizes of behaviors, we find that 95% of modifications are benign. Most of them are due to misclassification and misconfiguration. Interestingly, we observe that the behaviors of long-term proxies are quite consistent: they either always perform the content modification or never do it. No long-term proxies are found to intermittently modify the content.

Performance. The performance of long-term proxies and short-term proxies is shown in Table 16. The long-term proxies demonstrate clearly better performance than short-term proxies. Long-term proxies' RTT is about half of the short-term's proxies, and Long-term proxies' download speed is nearly four times of short-term proxies.

In this section, we compare long-term proxies with short-term proxies from different aspects. Our analysis shows that long-term proxies have better performance than short-term proxies. The reasons why long-term proxies can exist for a long time are (1) they are well managed by excellent hosting providers; (2) they are misclassified by proxy collectors for a long time, but proxy collectors falsely publish them. (3) owners accidentally misconfigured such proxies to be open to any user and owners does not notice that and remedy them.

8. Discussion

8.1. Ethical considerations

In this study, we collect open proxies from published open proxy lists. We do not utilize large scale port scanning to detect open proxies. Thus, the normal usage of open proxies is not affected, and private proxies are not exposed. In addition, open proxies are used to access our designed static websites that do not cause any harm to open proxies. The collected data does not include any open proxy owners' and other users' personal and private information. In summary, this study does not bring any risk and damage to proxy owners and users.

8.2. Limitations

We share similar approaches with earlier research to detect content modification, which cannot determine if the behavior-changing proxies have a hidden malicious purpose. The previous studies also have the same limitations. Our open proxy sources may not be complete, and some open proxies may not be included in our dataset. However, we have attempted to find as many open proxy lists as possible, which can be automatically crawled and downloaded to shorten the experiment time and enrich our proxy dataset. Moreover, in this study, we have also collected and tested most open proxies in the previous works.

We use the DOM tree structures to identify content modifications similar to Tsirantonakis's work [23]. However, in this work, we only use DOM structures to determine whether contents are modified and which parts are modified. To identify proxy groups that share similar behaviors, we employ a new approach that extracts owner information from elements, including metadata (title, keywords, and other fields), inject library, and URLs by parsing the HTML content. By combining the DOM structures and parsed owner information, we can quickly and accurately identify proxy owners and then group them.

We do not investigate whether the open proxies modify dynamic elements since it is challenging to decide whether the changes of dynamic elements are caused by themselves or open proxies. In the future, we will design websites that include dynamic elements to test open proxies and introduce new methods to distinguish the modifications caused by open proxies from those made by websites.

8.3. Comparisons with other studies

Here we present a comparison of our work with existing open proxy studies and highlight the improvement of our study from prior work.

In this study, we conducted a larger-scale analysis of the open proxy ecosystem. Table 18 lists the number of collected and responsive proxies in related studies and our work. Among those studies, the size of our collected open proxy dataset is the second-largest. Note that, although the study [27] examined a larger open proxy dataset, it lacks the active measurements and verification process of open proxies as it only analyzed open proxies based on the passive measurements. By contrast, our study combines active and passive measurements to investigate the open proxy ecosystem. Furthermore, as shown in Table 18, our study collects and examines significant more responsive open proxies than other studies, and those responsive open proxies are more critical and representative in the open proxy ecosystem.

Table 18 compares the percentage of identified content modifications and studied modification types. Our work presents the most comprehensive analysis on the misbehavior of open proxies. Table 17 shows the research content of the open proxy studies. We first analyzed cloud-based proxies and long-term proxies. In particular, although the work done by Scott et al. [24] examined several specific open proxy server owners, our study is the first to identify and analyze the groups of open proxy owners and their behaviors in a systematic manner.

9. Conclusion

This paper presents a comprehensive measurement study and in-depth analysis of the open proxy ecosystem. We conducted a large-scale measurement that collected more than 436 thousand proxies (including more than 104 thousand responsive proxies) over ten months. We characterized the open proxies' deployment, performance, and behaviors. We collected and analyzed large amounts of responses and classified open proxies based on their DOM tree structures. Furthermore, we identified and tracked the owners of open proxy groups by parsing HTML content and extracting identifier information. We analyzed the categories of content modification and deployment as well as the management strategy of malicious open proxies. We found that 76.42% of content modification proxies demonstrate malicious behaviors, among which Ad injection and redirection are the most prevalent activities. Our case studies show that malicious open proxy owners manipulate proxy deployment to increase their impacts by changing the deployment of their proxies (*e.g.*, the ASes and locations). Finally, we studied two specific groups of proxies, cloud-based proxies and long-term proxies. Our analysis shows that cloud-based proxies are a small portion of the open proxy ecosystem, but these proxies are more reliable and have better performance than non-cloud proxies. Meanwhile, long-term proxies demonstrate better performance than short-term proxies.

Table 17
Study content of open proxy studies.

Studies	Cloud-based proxies	Long-term proxies	Content manipulations	Owner study	Blacklist check
Scott [24]	×	×	×	✓	×
Tsirantonakis [23]	×	×	✓	×	✓
Perino [26]	×	×	✓	×	×
Mani [25]	×	×	✓	×	×
Choi [27]	×	×	×	×	✓
This study	✓	✓	✓	✓	✓

Table 18
The number of collected open proxies, the percentage of content modifications and modification types in existing studies.

Studies	# Collected	# Responsive	% of content modifications	Modification types
Scott [24]	4250	1880	N/A	N/A
Tsirantonakis [23]	65,871	19,473	5.15%	Tracking/Fingerprinting/Privacy leakage/Malware
Perino [26]	180,000	39,143	≈10%	Ad injection/Fingerprinting/Tracking
Mani [25]	107,034	31,000	≈8%	Ad injection/Cryptojacking/Eavesdropping/Malware
Choi [27]	1,045,468	N/A	N/A	N/A
This study	436,451	104,114	7.27%	Replacement/Ad injection/CSS injection/Redirection/Collect user information/Cryptojacking

CRediT authorship contribution statement

Rui Bian: Conceptualization, Methodology, Investigation, Software, Data curation, Writing – original draft, Visualization, Writing – review & editing. **Shuai Hao:** Methodology, Writing – review & editing. **Haining Wang:** Writing – review & editing, Supervision. **Chase Cotton:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

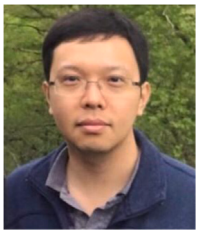
We would like to thank the anonymous reviewers for their insightful comments. This work is partially supported by Open Technology Fund (OTF) under an Internet Freedom Fund. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References

[1] T. Gerbet, A. Kumar, C. Lauradoux, A privacy analysis of google and yandex safe browsing, in: 2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), 2016, pp. 347–358.
[2] R. Kang, S. Brown, S. Kiesler, Why do people seek anonymity on the internet? Informing policy anddesign, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2013, pp. 2657–2666.
[3] A. McDonald, M. Bernhard, L. Valenta, B. VanderSloot, W. Scott, N. Sullivan, J.A. Halderman, R. Ensafi, 403 forbidden: A global view of cdn geoblocking, in: Proceedings of the Internet Measurement Conference 2018, 2018, pp. 218–230.
[4] Z. Weinberg, S. Cho, N. Christin, V. Sekar, P. Gill, How to catch when proxies lie: Verifying the physical locations ofnetwork proxies with active geolocation, in: Proceedings of the Internet Measurement Conference 2018, 2018, pp. 203–217.
[5] S. Burnett, N. Feamster, Encore: Lightweight measurement of web censorship with cross-originrequests, in: Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication, 2015, pp. 653–667.
[6] J.W. Byers, Encore: Lightweight Measurement of Web Censorship with Cross-Origin Requests–Public Review, Technical Report, 2015.
[7] Z. Weinberg, M. Sharif, J. Szurdi, N. Christin, Topics of controversy: An empirical analysis of web censorship lists, Proc. Priv. Enhanc. Technol. 2017 (1) (2017) 42–61.
[8] proxy-daily. https://proxy-daily.com/.
[9] proxylistdaily. https://www.proxylistdaily.net/.
[10] smallseotools. https://smallseotools.com/free-proxy-list/.
[11] dailyfreeproxy. https://www.dailyfreeproxy.com/.
[12] sinium. https://seopro.sinium.com/free-proxy-list.
[13] proxy-list.download. https://www.proxy-list.download/HTTP.
[14] openproxy.space. https://openproxy.space/list.
[15] proxyserverlist24. http://www.proxyserverlist24.top/.
[16] live-sock. http://www.live-socks.net/.
[17] ProxyBroker. https://github.com/constverum/ProxyBroker/.
[18] clarketm. https://github.com/clarketm/proxy-list.
[19] TheSpeedX. https://github.com/TheSpeedX/PROXY-List.
[20] opsxcq. https://github.com/opsxcq/proxy-list.
[21] fate0. https://github.com/opsxcq/proxy-list.
[22] a2u. https://github.com/a2u/free-proxy-list.
[23] G. Tsirantonakis, P. Ilia, S. Ioannidis, E. Athanasopoulos, M. Polychronakis, A large-scale analysis of content modification by open HTTP proxies, in: Network and Distributed System Security Symposium (NDSS), 2018.
[24] W. Scott, R. Bhoraskar, A. Krishnamurthy, Understanding open proxies in the wild, Chaos Commun. Camp (2015).
[25] A. Mani, T. Vaidya, D. Dworken, M. Sherr, An extensive evaluation of the internet's open proxies, in: Proceedings of the 34th Annual Computer Security Applications Conference, 2018, pp. 252–265.
[26] D. Perino, M. Varvello, C. Soriente, ProxyTorrent: Untangling the free HTTP (S) proxy ecosystem, in: Proceedings of the 2018 World Wide Web Conference, 2018, pp. 197–206.
[27] J. Choi, M. Abuhamad, A. Abusnaina, A. Anwar, S. Alshamrani, J. Park, D. Nyang, D. Mohaisen, Understanding the proxy ecosystem: A comparative analysis of residential and open proxies on the internet, IEEE Access 8 (2020) 111368–111380.
[28] K. Thomas, E. Bursztein, C. Grier, G. Ho, N. Jagpal, A. Kapravelos, D. McCoy, A. Nappa, V. Paxson, P. Pearce, et al., Ad injection at scale: Assessing deceptive advertisement modifications, in: 2015 IEEE Symposium on Security and Privacy, 2015, pp. 151–167.
[29] S.A. Alrwais, A. Gerber, C.W. Dunn, O. Spatscheck, M. Gupta, E. Osterweil, Dissecting ghost clicks: Ad fraud via misdirected human clicks, in: Proceedings of the 28th Annual Computer Security Applications Conference, 2012, pp. 21–30.
[30] S. Arshad, A. Kharraz, W. Robertson, Identifying extension-based ad injection via fine-grained web content provenance, in: International Symposium on Research in Attacks, Intrusions, and Defenses, 2016, pp. 415–436.
[31] R. Gomer, E.M. Rodrigues, N. Milic-Frayling, M. Schraefel, Network analysis of third party tracking: User exposure to tracking cookies through search, in: 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Vol. 1, 2013, pp. 549–556.
[32] L. Jin, S. Hao, H. Wang, C. Cotton, Your remnant tells secret: Residual resolution in ddos protection services, in: 2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), 2018, pp. 362–373.
[33] A. Fass, M. Backes, B. Stock, Hidenoseek: Camouflaging malicious javascript in benign asts, in: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, 2019, pp. 1899–1913.
[34] P. Skolka, C.-A. Staicu, M. Pradel, Anything to hide? Studying minified and obfuscated code in the web, in: The World Wide Web Conference, 2019, pp. 1735–1746.
[35] Y. Zhang, H. Zhang, X. Yuan, N.-F. Tzeng, Pseudo-honeypot: Toward efficient and scalable spam sniffer, in: 2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), 2019, pp. 435–446.
[36] V.S. Pai, L. Wang, K. Park, R. Pang, L. Peterson, The dark side of the web: an open proxy's view, ACM SIGCOMM Comput. Commun. Rev. 34 (1) (2004) 57–62.
[37] L. Wang, K. Park, R. Pang, V.S. Pai, L.L. Peterson, Reliability and security in the CoDeeN content distribution network, in: USENIX Annual Technical Conference (ATC), 2004, pp. 171–184.
[38] N. Weaver, C. Kreibich, M. Dam, V. Paxson, Here be web proxies, in: International Conference on Passive and Active Network Measurement, 2014, pp. 183–192.

[39] S. Huang, F. Cuadrado, S. Uhlig, Middleboxes in the internet: a HTTP perspective, in: Network Traffic Measurement and Analysis Conference (TMA), 2017, pp. 1–9.
[40] X. Mi, X. Feng, X. Liao, B. Liu, X. Wang, F. Qian, Z. Li, S. Alrwais, L. Sun, Y. Liu, Resident evil: Understanding residential ip proxy as a dark service, in: 2019 IEEE Symposium on Security and Privacy (SP), 2019, pp. 1185–1201.
[41] T. Chung, D. Choffnes, A. Mislove, Tunneling for transparency: A large-scale analysis of end-to-end violations in the internet, in: ACM Internet Measurement Conference (IMC), 2016, pp. 199–213.
[42] M. O'Neill, S. Ruoti, K. Seamons, D. Zappala, TLS proxies: Friend or foe? in: ACM Internet Measurement Conference (IMC), 2016, pp. 551–557.
[43] X.d.C. de Carnavalet, M. Mannan, Killed by proxy: Analyzing client-end TLS interception software, in: Network and Distributed System Security Symposium (NDSS), 2016.
[44] Z. Durumeric, Z. Ma, D. Springall, R. Barnes, N. Sullivan, E. Bursztein, M. Bailey, J.A. Halderman, V. Paxson, The security impact of HTTPS interception, in: Network and Distributed System Security Symposium (NDSS), 2017.
[45] G. Tyson, S. Huang, F. Cuadrado, I. Castro, V.C. Perta, A. Sathiaseelan, S. Uhlig, Exploring http header manipulation in-the-wild, in: International Conference on World Wide Web (WWW), 2017, pp. 451–458.
[46] G-Tools. https://github.com/jaxBCD/G-Tools.
[47] MaxMind GeoLite2 Databases. https://dev.maxmind.com/geoip/geoip2/geolite2/.
[48] Pydnsbl, Async dnsbl lists checker based on asyncio/aiodns. https://github.com/dmippolitov/pydnsbl.
[49] Norton community: repeated portscan issue with hn.kd.ny.adsl. https://community.norton.com/en/forums/repeated-portscan-issue-hnkdnyadsl.
[50] Blog: hn.kd.ny.adsl: Research, Ban. https://dontai.com/wp/2016/06/08/hn-kd-ny-adsl-research-ban/.
[51] Sophos community: Top hacker hn.kd.ny.adsl ?? https://community.sophos.com/utm-firewall/f/network-protection-firewall-nat-qos-ips/39664/top-hacker-hn-kd-ny-adsl.
[52] SUC012 : Chinese Blind SQL Injection – hn.kd.ny.adsl. https://eromang.zataz.com/2010/04/30/suc012-blind-sql-injection-china/.
[53] Spicework community: Hi been my network has been compppromised by a known port scammer, hn.kd.ny.adsl. https://community.spiceworks.com/topic/2301469-hi-been-my-network-has-been-compppromised-by-a-known-port-scammer-hn-kd-ny-adsl.
[54] Domain reputation: hn.kd.ny.adsl. https://talosintelligence.com/reputation_center/lookup?search=hn.kd.ny.adsl.
[55] otx.alienvault.com:azteca-comunicaciones.com. https://otx.alienvault.com/indicator/domain/azteca-comunicaciones.com.
[56] Live IP Map:200.69.79.170. https://www.liveipmap.com/200.69.79.170.
[57] Domain reputation: azteca-comunicaciones.com. https://talosintelligence.com/reputation_center/lookup?search=azteca-comunicaciones.com.
[58] Domain reputation: static.vnpt.vn. https://talosintelligence.com/reputation_center/lookup?search=static.vnpt.vn.
[59] The Spam Auditor Blog:SMTP AUTH Attacks, How Big is the Problem Really? https://spamauditor.org/2019/01/smtp-auth-attacks-how-big-is-the-problem/.
[60] abuseipdb:113.161.62.81. https://www.abuseipdb.com/check/113.161.62.81.
[61] vnpt.vn Content Scraper: Research, Ban. https://dontai.com/wp/2016/06/14/vnpt-vn-content-scraper-research-ban/.

Rui Bian is a Ph.D. student in the Department of Electrical and Computer Engineering at the University of Delaware. He received his B.Sc. and M.Sc. degrees in Engineering at University of Science and Technology of China in 2012 and 2015. His research interests include network security and web security.



Shuai Hao received his Ph.D. degree in Computer Science from the College of William and Mary, Williamsburg, VA, in 2017. He is an Assistant Professor in the Department of Computer Science at Old Dominion University (ODU), Norfolk, VA. Prior to joining ODU in 2019, he worked as a Postdoctoral researcher in CAIDA at UC San Diego. His research interests lie in the measurement and security of Internet infrastructure and networking systems.

Haining Wang received his Ph.D. in Computer Science and Engineering from the University of Michigan at Ann Arbor in 2003. He is a Professor in the Department of Electrical and Computer Engineering at Virginia Tech. His research interests lie in the areas of security, networking systems, cloud computing, and Internet-of-Things (IoT) systems. Before joining Virginia Tech in 2019, he was a Professor of ECE at the University of Delaware. He is a fellow of the IEEE.

Chase Cotton is currently a Professor of Electrical and Computer Engineering at the University of Delaware. His earlier research involved creating new methods in bridging, multicast, packet-based applications, traffic monitoring, transport protocols, custom VLSI, and Gigabit networking. He currently consults on communications and Internet architectures for many carriers and equipment vendors worldwide.