IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 6, 2020, accepted May 25, 2020, date of publication June 11, 2020, date of current version June 26, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3000959

Understanding the Proxy Ecosystem: A Comparative Analysis of Residential and Open Proxies on the Internet

JINCHUN CHOI[1,2], MOHAMMED ABUHAMAD[1,2], AHMED ABUSNAINA[2], (Graduate Student Member, IEEE), AFSAH ANWAR[2], (Graduate Student Member, IEEE), SULTAN ALSHAMRANI[2], JEMAN PARK[2], DAEHUN NYANG[3], AND DAVID MOHAISEN[2], (Senior Member, IEEE)

[1]Inha University, Incheon 22212, South Korea
[2]University of Central Florida, Orlando, FL 32816, USA
[3]Ewha Womans University, Seoul 03760, South Korea

Corresponding authors: Daehun Nyang (nyang@ewha.ac.kr) and David Mohaisen (mohaisen@ucf.edu)

This work was supported by the National Research Foundation under Grant NRF-2016K1A1A291275.

ABSTRACT Proxy servers act as an intermediary and a gateway between users and other servers on the Internet, and have many beneficial applications targeting the privacy of users, including bypassing server-side blocking, regional restrictions, etc. Despite the beneficial applications of proxies, they are also used by adversaries to hide their identity and to launch many attacks. As such, many websites restrict access from proxies, resulting in blacklists to filter out those proxies and to aid in their blocking. In this work, we explore the ecosystem of proxies by understanding their affinities and distributions comparatively. We compare residential and open proxies in various ways, including country-level and city-level analyses to highlight their geospatial distributions, similarities, and differences against a large number of blacklists and categories therein, *i.e.,* spam and maliciousness analysis, to understand their characteristics and attributes. We conclude that, while aiming to achieve the same goal, residential and open proxies still have distinct characteristics warranting considering them separately for the role they play in the larger Internet ecosystem. Moreover, we highlight the correlation of proxy locality distribution and five country-level characteristics, such as their Internet censorship, political stability, and Gross Domestic Product (GDP).

INDEX TERMS Residential proxy, open proxy, comparative analysis, geospatial analysis, blacklisting.

I. INTRODUCTION

Recently, a lot of efforts have been made to improve the privacy of users on the Internet, building an ecosystem around privacy enhancing infrastructure. Protecting user's privacy is an important concern in all areas of technology and business alike, and users utilize several approaches to protect their own privacy [1], [2]. For instance, proxy servers can be considered one of the easiest approaches for users to strengthening their privacy by hiding their actual Internet Protocol (IP) address [3], [4]. Proxy servers, shortly proxies, act as an intermediary for delivering online communication between users and Internet services (remote servers). By connecting to proxies, users do not have to directly send their request to the remote server (*e.g.,* web server) but to proxies. When a proxy receives a request from a user for a particular resource, the proxy first searches the internal cache for that resource and returns it to the user if found. If not found, the proxy forwards the request to the server to get a response, which is passed back to the user. The caching operation of proxies reduces the need for direct communication between users and remote servers, which leads to the prevention of network bottlenecks. Moreover, by sending and receiving packets through a proxy, users can avoid revealing their IP addresses to the remote servers.

Besides privacy protection, proxies can also be used to avoid Internet censorship. Users on the Internet may be censored by Internet providers and/or governments, in certain regions. Governments of various country can monitor their networks and block access to information and sites that are perceived as harmful (to the public or to the government). For citizens in those countries, a proxy can be an option to bypass governmental censorship and retrieve the information they seek. Rather than accessing a particular website directly,

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Omer Farooq.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

accessing it through a proxy in another country makes it less likely to be detected by the Internet censors [5].

A proxy that is open to the public is called an open proxy. Without any permission from the operator, users can utilize open proxies to protect their privacy and to access information that is otherwise restricted by local entities. The list of available open proxies is continuously updated and broadly posted on many websites [6], [7]. This accessibility often results in having the open proxies blacklisted easily. Furthermore, most open proxies have data center IP addresses, so web service providers can easily recognize whether a request is coming from a proxy or directly from a user [8].

On the other hand, while open proxies allow the users to hide their IP addresses and protect their privacy, a compromised proxy (or a rogue) can perform a malicious activity, e.g., between the user and the server. In particular, when the end-to-end encryption is not used, the malicious proxy can manipulate the contents of the transferred data or capture confidential information that is meant only for the user. Since operators and policies for open proxies in many cases are not well-defined, the security threats and implications for users using such open proxies can be significant [9]–[11].

Another type is residential proxy, where providers utilize IP addresses that are assigned by a general Internet Service Provider (ISP) for their use, which makes the request from a proxy looks more discreet. In general, the residential proxies are operated in closed fashion and only paid users are allowed to use a group of proxies owned by the operator. Both open proxies and residential proxies share similar characteristics, although residential proxies are different in how they are managed, *i.e.,* they are "generally" closed.

Motivation. Proxies contribute to improving the privacy of users on the Internet, while often being targeted for malicious behavior, which motivates our research. Given the different operation settings of open proxies compared to residential proxies, their distribution, regional background, and behaviors can be an characterization to understanding the proxy ecosystem, and their role in network security. For example, the usage of proxies is likely to be a result of regional policies and characteristics, and analyzing them can contribute to understanding the correlation between several aspects of regional-level characteristics and attributes. In this work, we analyze the geospatial distribution of proxies, both open and residential, at the country- and regional-level, to show characteristics related to location affinities and gain insights on their correlation with different country-level policies and attributes. We highlight distribution of blacklisted proxies and their correlation to countries policies, performance, and Internet speed. Using 27 blacklisting services, we highlight the variety of malicious activities of blacklisted proxies. We also provide a correlation analysis of proxies geospatial distribution and five country-level characteristics: Internet content censorship, Internet freedom, political stability, Internet speed, and gross domestic product. Our analysis shows that 79.11% of the open proxies are prone to blacklisting. Similarly, 86.04% of the residential proxies are prone to blacklisting. Moreover, we investigated the behavior of the proxies and found that 28.23% and 16.85% of the open and residential proxies were used for spam, respectively. In addition, 6.97% of the open proxies are associated with verified attacks, along with 0.27% of the residential proxies.

Contribution. Our main contributions are as follows:

- We investigate the geolocation distribution of a large dataset that includes 1,045,468 open proxies and 6,419,987 residential proxies. The locality distribution of proxies is conducted on the country-level, city-level, and autonomous system-level where the proxies reside.
- We analyze the behavior of the proxies using 27 different blacklisting services. We show that the majority of proxies are blacklisted, and 28.23% and 16.85% of open and residential proxies are used for spam, respectively. Moreover, we investigate the proxies that are associated with verified attacks. Our analysis shows that 6.97% of open proxies, along with 0.27% of residential proxies participated in launching malicious attacks.
- We conduct correlation analyses of proxies locality distribution and five country-level characteristics, showing a strong positive correlation between Internet speed and Gross Domestic Product GDP) with numbers of proxies within countries.

Organization. The rest of the paper is organized as follows: In section II we highlight the efforts toward understanding and analyzing the behavior of the Internet proxies. We describe the dataset used in this study, the preprocessing, geolocation distribution of the proxies, and their behavior in section III. In section IV, we conduct a correlation study to understand the relationship between the distribution of the proxies and five factors, including the censorship, Internet freedom, political stability, Internet speed, and the gross domestic product. Finally, we conclude our work in section V.

II. RELATED WORK

Recently, several studies have been exploring the ecosystem of proxies by analyzing their behavior and performance, as well as the security aspects of such services [12]. While most of the studies addressed different aspects related to open proxies, few works have been done toward analyzing residential proxies due to the challenges in identifying them. Addressing and analyzing the distribution of both open and residential proxies and their relation to regional characteristics is the main goal of this study which fills the gap in current literature. This section highlights the efforts towards understanding and analyzing the behavior of proxies.

Open Proxy. To fully-understand the reliability and the security of open proxies, Mani *et al.* [13] have conducted a comprehensive study on open proxies using a large-scale dataset of 107,000 listed open proxies and 13 million proxy requests over a 50-day period. The authors concluded that 92% of the listed open proxies are unresponsive to proxy requests. Further, the study also found that a substantial number of open proxies have a sort of malicious behavior, *e.g.,* modifying the Hypertext Markup Language (HTML) content to be used for cryptocurrency mining (cryptojacking), launching man-in-the-middle attacks, fetching remote access Trojans and/or other forms of malware. Tsirantonakis *et al.* [14] proposed a framework that collects Hypertext Transfer Protocol (HTTP) proxies from different websites, and tests them using decoy websites-based methods (dubbed honeysites). The study implemented a content modification detection technique that aims to detect any object modifications by operating at the level of the page's Document Object Model (DOM) tree. Applying this technique on a dataset of (19,473) open proxies, the authors reported that 5.15% of the proxies perform a malicious content modification or injection. They also reported that 47% of the malicious proxies inject ads, 39% inject script to collect user data, and 12% used to redirect the user to malicious websites that contain malware. Even with such risks, Perino *et al.* [15] showed that open proxies services are increasing drastically, and only a small fraction of the available proxies actually works. In their study, Perino *et al.* [15] reported that around 10% of the working proxies have a sort of malicious behavior.

Another work by Chung *et al.* [16] studied the end-to-end connectivity violation of the proxy services, where they utilized Luminati to detect end-to-end violations of Domain Name Server (DNS), HTTP, and Hypertext Transfer Protocol Secure (HTTPS), and to detect when a host or an ISP perform a content monitoring. Using more than 1.2 million nodes across 14,000 autonomous systems covering 172 countries, the findings showed that 4.8% of nodes are subject to some type of end-to-end connectivity violation. The reliability of proxies can be measured by how the advertised location is accurate. Recent studies such as Weinberg *et al.* [17] have shown that some proxies providers are advertising to have a wide range of locations, while in fact their proxies are in certain countries in which the server cost is cheap. Another work by Weaver *et al.* [18] utilized Netalyzr and techniques based on traceroutes of the responses to TCP connection to detect the presence of proxies.

Residential Proxy. The first study examining the behavior of the residential proxies is due to Mi *et al.* [8], where the authors conducted an in-depth analysis on residential proxy services and servers, including about six million residential IP addresses across ≈230 countries and 52,000 Internet Service Providers (ISPs). Their findings show that even though residential proxy providers claim that the proxy hosts willingly participated in providing the service, many proxies operate on compromised hosts. They also reported Potentially Unwanted Programs (PUP) logs as well as other malicious activities, such as ads, phishing, and malware hosting.

III. DATA COLLECTION AND MEASUREMENT

A. PROXY DATA COLLECTION

For open proxies, we used the dataset provided by IP2Proxy [19], which makes up a large portion of our dataset. We also searched websites listing open proxies, and regularly collected the proxy IP addresses from them as of November 2019. Residential proxies are not public, so it is difficult to obtain their IP addresses in a similar way. To this end, we obtained the dataset residential proxies from Mi *et al.* [8]. Mi *et al.* utilized an infiltration framework to collect a dataset of 6,419,987 residential proxies distributed across more than 230 countries and more than 52,000 ISPs. The captured IPv4 addresses acting as residential proxies were observed using five residential proxy providers between July 2017 and March 2018. Figure 1(b) shows the locality distribution of the residential proxies, and Table 1 shows the country/region distribution of the top 10 localities for those proxies.

TABLE 1. Country/region-level distribution of open and residential proxies. China and the USA contain approximately 29% of the open proxies, while they are not in the top 10 countries/regions in the residential proxies list. Similarly, Turkey contains 528,032 residential proxies, but only 5,040 open proxy.

Proxy List	# Proxies
IP2PROXY [17]	1,041,455
MultiProxy [18]	2,230
clarketm [19]	1,500
checkerproxy.net [20]	50,190
proxybroker [21]	5,441
Total (unique)	1,045,468

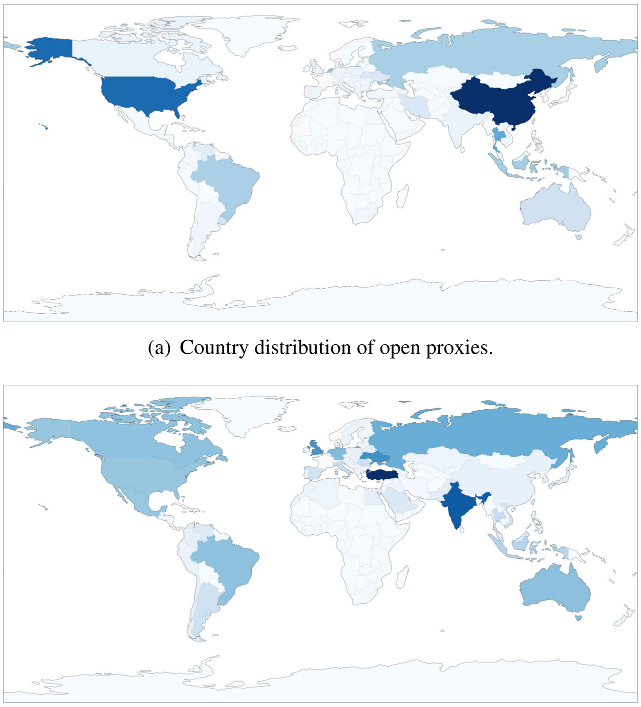
Open Proxy and Residential Proxy. We collect open proxy IP addresses by searching Google for an "open proxy list" and selecting sites that are updated regularly or can perform direct collection commands. We summarize the collected number of open proxy IP addresses in Table 2. A large proportion of the dataset is from IP2Proxy [19], with a total number of unique IP addresses of 1,045,468. We observe that different lists provide same IP addresses (55,348 IP addresses). To gather residential proxies IP addresses, we obtained residential proxy dataset from Mi *et al.* [8]. This dataset consists of IPv4 addresses collected between July 2017 and March 2018 and contains a total of 6,419,987 IP addresses. We find that there are common IP addresses between the two datasets. That is, 20,816 IP addresses exist in both open and residential proxy datasets. After collecting the datasets, we conduct a geospatial analysis to obtain the distribution of open and residential proxies. We categorize the locations of proxies by country-, city-, and autonomous system-level locality of the proxies. We start by obtaining the geolocation and Autonomous System Number (ASN) of each IP address using the IP-to-region local dataset and the MaxMind online database [24].

B. GEOSPATIAL ANALYSIS

Country-level Distribution of Proxies. Figure 1 shows the country distribution of open and residential proxies. The darker blue shade indicates a higher number of proxies in the given country. Figure 1(a) describes the city-level distribution of open proxies with China and the US accounting for a large proportion, as they occupy 28.7% of all open proxies. The distribution of residential proxy is shown in Figure 1(b),

TABLE 2. Websites that provide open proxy lists and the number of open proxy IP addresses collected. We can see that there are many duplicate IP addresses collected.

Rank	Open Proxy			Residential Proxy		
	Country/Region	# Proxies	% Proxies	Country/Region	# Proxies	% Proxies
1	China	169,431	16.21%	Turkey	528,032	8.22%
2	USA	131,302	12.56%	India	440,215	6.86%
3	Thailand	88,624	8.48%	Ukraine	331,091	5.16%
4	Netherlands	68,506	6.55%	UK	320,375	4.99%
5	Indonesia	60,140	5.75%	Russia	264,863	4.13%
6	Russia	57,675	5.52%	Germany	234,291	3.65%
7	Brazil	57,031	5.46%	Netherlands	228,707	3.56%
8	Australia	33,903	3.24%	Australia	221,853	3.46%
9	Taiwan	27,609	2.64%	Canada	217,633	3.39%
10	Uruguay	27,330	2.61%	Brazil	216,989	3.38%
Total	Worldwide	1,045,468	100%	Worldwide	6,419,987	100%



(a) Country distribution of open proxies.

(b) Country distribution of residential proxies.

FIGURE 1. Country distribution of open and residential proxies. Darker shade of blue represents more proxies residing in the country. Here, China and the US contain the majority of Open Proxies, while Turkey and India contain the highest number of residential proxies.

which is different from the distribution of the open proxy. Turkey and India have a large portion (15.08%), followed by Ukraine and the United Kingdom. Table 1 provides the top 10 country/region distributions of open and residential proxies in our data collection. The distribution of open proxies by country is concentrated in the top two countries, but the residential proxies are more dispersed in Russia and European countries and South America. The top 10 nations of the open proxy account for nearly 70%, while residential proxy accounts for less than 50% (46.8%).

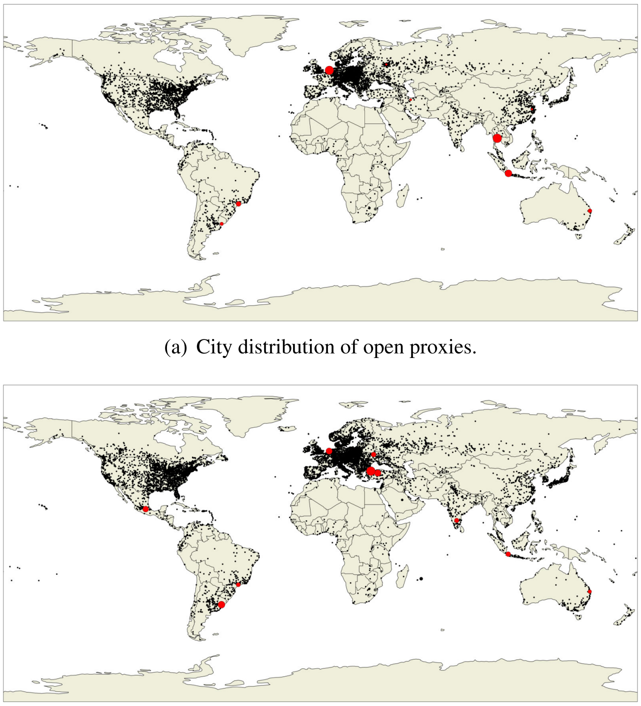
City-level Distribution of Proxies. The distribution of cities in open and residential proxies is similar to that of countries. However, China is located at the top of the country distribution. We find only one Chinese city (Hangzhou) in the top 10 ccities, as shown in Table 3. This indicates that the proxy is scattered in many cities in China, where nearly 300 Chinese cities appear in our dataset. Figure 2(a) describes city distribution of open proxies. In this figure, we use circles to present the number of open and residential proxies in each city. Also, we highlight the top 10 cities with the red color and larger size. The size of the circle depends on the number of open proxies in the city. To better illustrate the distribution within the cities, the region should be limited to a specific country. Figure 3(a) shows the distribution of open proxies in China. This figure shows that not only Hangzhou but also other Chinese cities occupy a large number of open proxies. The large 10 circles in this figure represent the top 10 cities in China with open proxy numbers, which are in the top 30 cities of the entire open proxy. This indicates that the open proxies in China are distributed among major cities, such as Hangzhou, Nanchang, Nanjing, Guangzhou, and Beijing.

Another example concerns the United States, which has a high percentage of open proxy. It is ranked second in the open proxy distribution at the country-level, but in the city distribution, no city appears in the top 10. This means that the open proxy in the United States is evenly distributed in many cities. Figure 3(b) shows the distribution of open proxies in cities in the United States. There are a larger number of open proxies, distributed throughout the region and especially in densely populated areas in the east and west. Despite a large number of open proxies in the United States, only three cities were included in the top 30, ranked as 17th, 23rd, and 27th.

In the case of the residential proxy, the distribution of the city-level is more interesting. Figure 2(b) presents the distribution of urban levels of residential proxies. The cities of the Netherlands and other cities in European countries are similarly distributed, as shown in Figure 2(a). On the other hand, the two Turkish cities, Istanbul and Ankara, had an

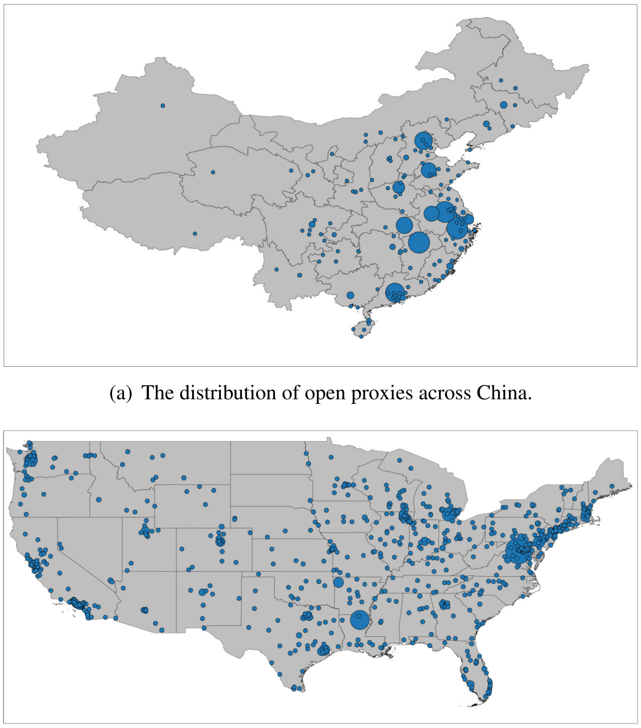
TABLE 3. City-level distribution of open and residential proxies. Bangkok and Amsterdam contain approximately 13% of the open proxies. While China is ranked first in the number of open proxies, only one city (Hangzhou) is in the top 10, indicating the high distribution of proxies across the country. Most of the residential proxies within Turkey are residing in Istanbul and Ankara (88.66%), as they are ranked first and third.

Rank	Open Proxy			Residential Proxy		
	City	# Proxies	% Proxies	City	# Proxies	% Proxies
1	Bangkok	68,094	6.51%	Istanbul	266,390	4.15%
2	Amsterdam	66,908	6.40%	Montevideo	211,860	3.30%
3	Jakarta	54,476	5.21%	Ankara	201,771	3.14%
4	Sao Paulo	38,617	3.69%	Amsterdam	192,072	2.99%
5	Brisbane	30,894	2.96%	Mexico City	180,613	2.81%
6	Montevideo	27,330	2.61%	Kiev	151,372	2.36%
7	Nonthaburi	20,402	1.95%	Jakarta	147,214	2.29%
8	Tehran	18,036	1.73%	Bangalore	138,312	2.15%
9	Hangzhou	17,631	1.69%	Sao Paulo	130,247	2.03%
10	Moscow	17,532	1.68%	Brisbane	113,339	1.77%
Total	Worldwide	1,045,468	100%	Worldwide	6,419,987	100%


(a) City distribution of open proxies.

(b) City distribution of residential proxies.

FIGURE 2. City distribution of open and residential proxies. The circle size reflects the number of proxies. In general, open and residential proxies and evenly distributed across all Europe, particularly in Ankara and Istanbul.


(a) The distribution of open proxies across China.

(b) The distribution of open proxies across the United States.

FIGURE 3. The city-level distribution of open proxies in China and the United States. The circle size reflects the number of proxies.

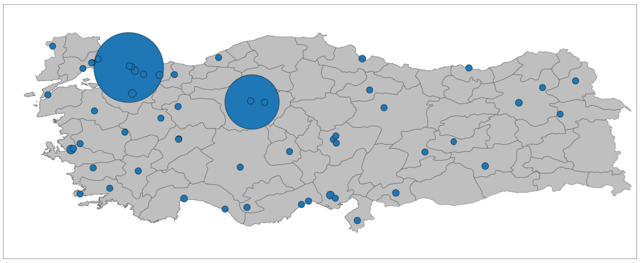
inconspicuous number of open proxies, as they were ranked first and third in the top 10 of the number of the residential proxy, respectively. In order to learn more about Turkey's residential proxy distribution, we represent the distribution of proxies within Turkey, shown in Figure 4(a). As mentioned earlier, the two Turkish cities, Istanbul and Ankara, have a large share. This may be due to the fact that almost 90% of Turkey's population lives in two cities.

We notice that India has the second largest number of residential proxies at the country-level, with four cities from India being in the top 30 of the residential proxy city-level distribution, and 2.15% of residential proxies of the world are located in Bangalore. Figure 4(b) shows the city-level residential proxy distribution in India. The four cities mentioned above are represented by large circles, and the other cities are widely distributed.

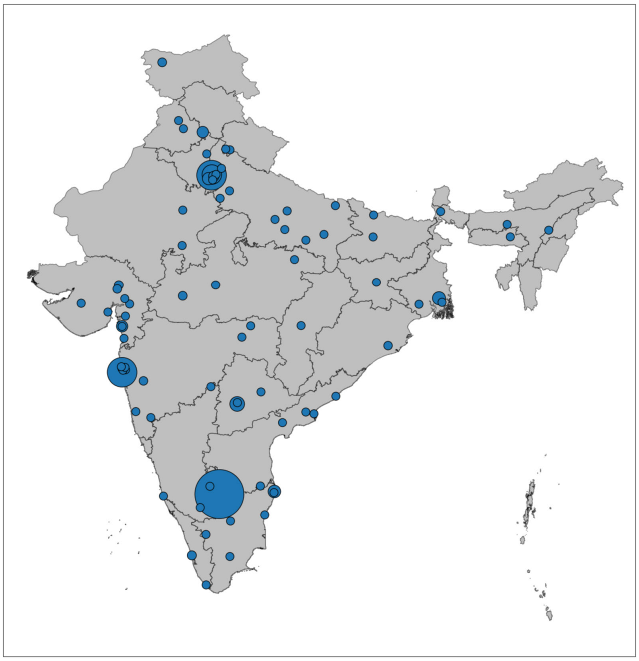
Distribution of proxies over ASs. We also analyzed the ASNs containing the IPs of the open and residential proxy and summarized them as shown in Table 4. From this analysis, we noticed that ASN 4134, which has the largest share of

TABLE 4. AS-level distribution of open and residential proxies. Here, ASN 4134 is covering China, and contains 10.92% of the open proxies. ASN 23969 is in Thailand, and contains 4.43% of the open proxies. Similarly, ASN 47331 is in Turkey, and cover 3.67% of the residential proxies.

Rank	Open Proxy			Residential Proxy		
	ASN	# Proxies	% Proxies	ASN	# Proxies	% Proxies
1	4134	114,116	10.92%	47331	235,474	3.67%
2	23969	46,304	4.43%	8151	154,318	2.40%
3	7713	39,158	3.75%	9829	150,951	2.35%
4	209	30,599	2.93%	9121	119,218	1.86%
5	4837	25,284	2.42%	7713	102,055	1.59%
6	3462	25,041	2.40%	3320	98,045	1.53%
7	45758	19,802	1.89%	24560	94,237	1.47%
8	14061	18,898	1.81%	25019	81,128	1.26%
9	8048	15,379	1.47%	2856	78,754	1.23%
10	131090	14,360	1.37%	12389	75,589	1.18%
Total	Worldwide	1,045,468	100.00%	Worldwide	6,419,987	100.00%



(a) City distribution of Residential Proxies in Turkey.



(b) City distribution of Residential Proxies in India.

FIGURE 4. City distribution of residential proxies in Turkey and India. The circle size reflects the number of proxies.

open proxy, serves China. We also notice that China has the largest number of open proxies. It is worth noting that ASN 4837, which accounts for 2.42% of the open proxy, is also an AS in charge of China. ASN 23969, ASN 45758 and ASN 131090, which are responsible for Thailand, account for 4.43%, 1.89% and 1.37%, respectively, summing to 80,466 proxies, representing 91% of Thailand's open proxy. This indicates that three ASes service most of Thailand's open proxy. In the AS-level distribution of residential proxies, ASN 47331 and ASN 9121 serve Turkey and they account for 3.67% and 1.86% of the total residential proxies, respectively. As noted earlier, Turkey has the largest number of residential proxies. The ASes that cover India in the AS-level distribution are ASN 9829 and ASN 24560, which serve more than 50% of the residential proxy located in India.

C. BLACKLISTS AND MALICIOUS BEHAVIOR

Proxies can be used by users to hide their identities. Although they are important for privacy assurance, proxies can also be a challenge to web security and administrators. It is necessary for the administrators to employ access control to their servers by knowing their customers and defend against fraudulent access [25]. Common methods for access control include manual and automated solutions. Manual blocking requires understanding the types of proxies, along with maintaining an updated list of proxy IP addresses.[1]

With this in mind, we attempt to identify the blacklisted IP addresses. To do so, we begin by collecting a list of blacklisting services. In total, we assemble a list of 27 such services. Leveraging those services, we then distribute the blacklisted proxies based on their intent, such as spammer, zombie risk, probable spammer, etc. Additionally, we argue that a proxy IP address if involved in malicious activities will be blacklisted and their intent identified at some point in time. However, it is known for a proxy IP address to be dynamic, meaning that an IP involved in an attack today may be assigned to a harmless service. Considering this, the blacklisting services allow a service to appeal against its IP address being blacklisted. Taking these into consideration, we aim to understand the distribution, patterns, and associations among them. For this study, we limit ourselves to the categories that

[1] Proxy IP addresses change on daily basis.

strictly identify a proxy to have been involved in spamming or attacks. In this section, we describe the different blacklisting services leveraged and how we distribute them into classes for further analysis.

Blacklist Services. To allow users to identify the different intents of proxies, there are multiple online services that make their list of blacklisted proxies public and classify their IP addresses depending on the posed challenge to a destination web-service. For example, the Real-time blockhole list *all.spam-rbl.fr* classifies proxies into spammer, zombie risk, etc. Additionally, these services frequently update their lists, e.g., *all.spam-rbl.fr* updates its list 10 times in a day.

- **Realtime Blackhole List (RBL).** RBL maintains lists of IP addresses that are susceptible to be used for spam. It maintains many lists of such IP addresses, depending on the source. We utilize the list that stores all the IP addresses listed, and later identify the intent based upon the return code by their API against our request for that proxy.
- **Spamrats.** Spamrats maintains multiple APIs based on the intent of the source. Each of these APIs maintain a set of blacklisted IPs. Among these APIs, we utilize the ones that store a set of IP addresses that are shown to be involved in spamming attacks or AUTH attacks. In AUTH attacks, a malicious user tries credentials obtained from breaches to authenticate. It particularly targets users that re-use their credentials across different services.
- **Weighted Private Block List (WPBL).** WPBL passively detects spams, with no crowd-sourced or manual additions. Additionally, they suggest securing the host and fixing misconfigurations to eliminate spam, and also provide a lookup facility to help users de-list themselves.
- **Uceprotect.** Uceprotect maintains APIs that list IP addresses with either wrong or missing or generic reverse DNS (PTR record), or dialup connections (typically suggesting a home/other user with a dynamic connection), or computers with exploited / exploitable security holes (e.g., open proxies, open relays, vulnerable webservers, virus infected, etc.) or which are assigned to well-known spammers. We limit ourselves to the proxies that are known to be spam sources by the service.
- **Justspam.** Justspam checks if an IP is listed by other well-known or independent blacklisting services. They claim to be a safeway to prevent false positives.
- **Sorbs.net.** sorbs.net maintains multiple APIs with lists of IPs by their intent, such as open HTTP proxy servers, IPs with spammer abusable vulnerabilities, known spam sources (last 48 hours/28 days/one year/anytime), hijacked, etc. It also lists spam supporting service providers with "third strike and you are out" basis. We limit ourselves to the lists that include spam and attack sources.
- **Junkemailfilter.** Junkemailfilter maintains lists of blacklist, yellowlist, brownlist, and whitelist IP addresses. We limit ourselves to the blacklist.
- **Korea services.** This service lists most IP address ranges (network address) assigned to Korea by APNIC, and any older ARIN ranges with a history of spam.
- **Spamhaus.** This popular service lists verified spammers, Register of Known Spam Operations (ROKSO), illegal third-party exploits, worms, and trojan horses.
- **DBUDB.com.** IPs are added to the DBUDB.com database automatically with no provision of manual addition. Addition to this list occurs when the recorded events for a given IPv4 address indicate substantially that a message content was spam, scam, virus, or other malware. IPs are added within 10 minutes or less of an outbreak; data is collected in real-time and the zone is updated every 10 minutes.

Limitations. The residential proxies dataset was collected between July 2017 and March 2018, while our blacklisting analysis is done in 2019. Such an observation/analysis time difference could introduce some false alarms on the number of blacklisted residential proxies since an IP address could be associated with a residential proxy for a specific period of time (*e.g.,* during the data observation/collection time) and then being associated with malicious activities later (*e.g.,* during the blacklisting analysis). In this study, the reported results do not take such a scenario into consideration given the limitations in investigating the period when the IPs acted as a residential proxy and the lack of information by the blacklisting services on the date in which an IP address was added to a certain blacklist. Given the large-scale dataset of proxies used in this study, consisting of 1,045,468 open proxies and 6,419,987 residential proxies (a total of 7,465,455 proxies), the impact of such limitation becomes less obvious, and therefore the analysis provides insights into the general behavior of the proxy ecosystem.

Country-level Analysis. Leveraging the blacklist services, we check if an open or residential proxy is present in any of the above blacklists. Among them, we then check if it is a proven spam, or if it shown to be involved in an attack, and if it has a vulnerability that can be exploited for future spam activities. Table 5 shows the results of the open proxies analysis. We observe that China has the highest number of IPs included in the blacklisting services, i.e., 94.24% of all the open proxies in the country. Additionally, it also has the highest number of proxies shown to be involved in spam activities and attack sources around the globe, and is the second country by the number of vulnerable sources. However, it has less than one percent vulnerable proxies. On the other hand, Iran stands at number 10 among the most blacklisted source-countries with ≈93% of its open proxies blacklisted, but is at the sixth position in the countries involved in attack sources and vulnerable sources. Other noteworthy countries and regions are Thailand and Taiwan, with almost 99.5% and 98% of their open proxies blacklisted, respectively. Conversely, the USA,

TABLE 5. Country/region-level distribution of blacklisted open proxies. The number of blacklisted proxies is proportional to the total number of proxies within the country/region. As shown, 99.42% of the proxies in Thailand are blacklisted. China and Thailand contain 29.96% of the blacklisted open proxies worldwide.

Rank	Country/Region	Blacklisted		Spam		Attack		Vulnerable	
		#	%	#	%	#	%	#	%
1	China	159,681	94.24%	67,040	39.57%	28,495	16.82%	1,613	0.95%
2	Thailand	88,115	99.42%	23,592	26.62%	8,844	9.98%	666	0.75%
3	USA	72,879	55.50%	18,955	14.44%	3,952	3.01%	726	0.55%
4	Indonesia	54,191	90.11%	14,318	23.81%	1,505	2.50%	3,121	5.19%
5	Brazil	47,671	83.58%	16,093	28.22%	4,743	8.32%	1,306	2.29%
6	Russia	44,277	76.77%	15,408	26.72%	2,705	4.69%	1,403	2.43%
7	Netherlands	43,310	63.22%	22,037	32.17%	3,913	5.71%	1,201	1.75%
8	Taiwan	27,160	98.37%	5,174	18.74%	351	1.27%	94	0.34%
9	Australia	26,402	77.87%	13,446	39.66%	2,656	7.83%	1,574	4.64%
10	Iran	23,189	92.98%	10,940	43.87%	2,999	12.02%	1,218	4.88%
Total	Worldwide	827,106	79.11%	295,152	28.23%	72,914	6.97%	21,035	2.01%

TABLE 6. Country-level distribution of blacklisted residential proxies. The number of blacklisted proxies is proportional to the total number of proxies within the country. Turkey contains 9.05% of the blacklisted proxies, with a blacklisting rate of 97.68%.

Rank	Country/Region	Blacklisted		Spam		Attack		Vulnerable	
		#	%	#	%	#	%	#	%
1	Turkey	515,767	97.68%	39,424	7.47%	191	0.04%	4,204	0.80%
2	Indonesia	432,780	98.31%	152,261	34.59%	603	0.14%	30,025	6.82%
3	UK	289,049	90.22%	11,608	3.62%	420	0.13%	635	0.20%
4	Ukraine	271,088	81.88%	33,759	10.20%	330	0.10%	5,289	1.60%
5	Russia	239,733	90.51%	47,501	17.93%	522	0.20%	7,229	2.73%
6	Germany	224,281	95.73%	17,365	7.41%	74	0.03%	1,035	0.44%
7	Mexico	203,584	99.29%	27,650	13.49%	60	0.03%	3,668	1.79%
8	Brazil	198,764	91.60%	19,571	9.02%	743	0.34%	3,539	1.63%
9	Uruguay	190,228	89.79%	25,587	12.08%	792	0.37%	3,608	1.70%
10	Australia	175,270	79.00%	45,465	20.49%	822	0.37%	11,965	5.39%
Total	Worldwide	5,700,244	86.04%	1,081,779	16.85%	17,596	0.27%	165,328	2.58%

although is at the third position in the number of blacklisted open proxies, it only represents 55.5% of its open proxies, which makes it the least blacklisted country by the percent representation. On the other hand, the analysis of residential IP addresses in Table 6 reveal that every country (except Ukraine and Australia) in the top 10 countries with highest number of residential IP addresses have more than 90% of their IPs blacklisted by one or more of the services, with Turkey, Indonesia, Germany, and Mexico having more than 95% blockage.

We also observe four countries—Indonesia, Russia, Brazil, and Austalia—in the top 10 blocked open and residential proxies. We observed that 99.3% of residential IP addresses in Mexico being blacklisted by at least one blacklisting services, and with 13.5% of its IP addresses being flagged for spam activities, 0.03% for launching attacks, and 1.6% for being vulnerable to future spam activities. Additionally, Indonesia and Australia are the countries that have most vulnerabilities that may lead to their involvement in spam activities in the future. Moreover, India, Vietnam, and Korea are among the top three countries with highest number of residential IPs with proven spam activities. Thailand, Vietnam, and Mauritius are the top three countries with highest representation of proven attacks, and India, Indonesia, and Australia represent the top three countries with most vulnerable IP addresses. India has the most residential IPs that have been involved in spam and highest number of vulnerable IPs that can be exploited for spam activities in the future.

City-level Analysis. Table 7 shows the top 10 cities with blacklisted open proxies, according to our analysis. While Bangkok appears as the city with highest number of blacklisted open proxies, Bangkok, Nonthaburi, Hangzhou, Nanchang, and Nanjing all have more than 99% of their open proxies blacklisted. However, only Nanjing has over 90% of its proxies involved in proven spam activity. Additionally, although 99.3% of the open proxies in Bangkok are blacklisted by the aforementioned services, only 28.74% of its proxies are proven to carry out spam activities.

TABLE 7. City-level distribution of blacklisted open proxies. The number of blacklisted proxies is proportional to the total number of proxies within the city. Bangkok contains 8.18% of the blacklisted open proxies, with a blacklisting rate of 99.30%. Note that most of the cities in this list are with a blacklisting rate of higher than 99%.

Rank	City	Blacklisted		Spam		Attack		Vulnerable	
		#	%	#	%	#	%	#	%
1	Bangkok	67,620	99.30%	19,572	28.74%	7,790	11.44%	595	0.87%
2	Jakarta	49,607	91.06%	12,171	22.34%	1,192	2.19%	2,640	4.85%
3	Amsterdam	42,477	63.49%	21,694	32.42%	3,839	5.74%	1,190	1.78%
4	Sao Paulo	32,364	83.81%	11,536	29.87%	3,314	8.58%	887	2.30%
5	Brisbane	24,734	80.06%	12,847	41.58%	2,550	8.25%	1,536	4.97%
6	Montevideo	21,687	79.35%	5,468	20.01%	1,299	4.75%	424	1.55%
7	Nonthaburi	20,396	99.97%	3,963	19.42%	1,043	5.11%	54	0.26%
8	Hangzhou	17,578	99.70%	8,288	47.01%	4,154	23.56%	169	0.96%
9	Nanchang	17,330	99.98%	4,400	25.39%	2,377	13.71%	108	0.62%
10	Nanjing	16,904	99.86%	15,270	90.21%	6,393	37.77%	289	1.71%
Total	Worldwide	827,106	79.11%	295,152	28.23%	72,914	6.97%	21,035	2.01%

TABLE 8. City-level distribution of blacklisted residential proxies. The number of blacklisted proxies is proportional to the total number of proxies within the city. Both Istanbul and Ankara are at the top of the list with a blacklisting rate of more than 97%. As shown, Mexico City is ranked fourth with 99.75% of the residential proxies blacklisted.

Rank	City	Blacklisted		Spam		Attack		Vulnerable	
		#	%	#	%	#	%	#	%
1	Istanbul	259,481	97.41%	21,665	8.13%	94	0.04%	2,381	0.89%
2	Ankara	197,607	97.94%	12,767	6.33%	79	0.04%	1,015	0.50%
3	Montevideo	190,228	89.79%	25,587	12.08%	792	0.37%	3,608	1.70%
4	Mexico City	180,158	99.75%	25,798	14.28%	47	0.03%	3,343	1.85%
5	Jakarta	145,197	98.63%	41,965	28.51%	251	0.17%	13,153	8.93%
6	Amsterdam	144,624	75.30%	40,241	20.95%	917	0.48%	7,096	3.69%
7	Bangalore	138,194	99.91%	56,800	41.07%	151	0.11%	9,855	7.13%
8	Kiev	130,989	86.53%	11,024	7.28%	125	0.08%	1,057	0.70%
9	Sao Paulo	117,910	90.53%	12,468	9.57%	543	0.42%	2,317	1.78%
10	Brisbane	105,951	93.48%	43,865	38.70%	809	0.71%	11,838	10.44%
Total	Worldwide	5,700,244	86.04%	1,081,779	16.85%	17,596	0.27%	165,328	2.58%

Moreover, Hangzhou has the most number (23.56%) of its open proxies involved in attacks and around 44% of its proxies are involved in spam activities. On the other hand, Table 8 shows the cities with most blacklisted residential IPs. It can be observed that all (except for Amsterdam and Kiev) have more than 90% of their residential IPs blacklisted, with four of them having more than 95% of their blacklisted. Interestingly, Bengaluru, India has 99.91% of its residential IP addresses blacklisted by one or more of the services, and more than 41% (highest by cities) of them are proven to be used for spam activities and more than 7% of the city's residential IPs vulnerable to future spam campaigns. Additionally, Brisbane has more than 10.4% of its residential IPs vulnerable and around 39% of its IPs proven spammers. The vulnerable residential IPs, if exploited, could make Brisbane the next most spam source-city in the world.

ASN-level Analysis. Similarly, as in Table 9 ASNs, e.g., 4134, have all their open proxies blacklisted but only 47.66% of them are proven to be involved in spam activities and 21.95% of them are involved in attacks. This can be because blacklisting services, such as uceprotect, blacklist all the IP addresses corresponding to the worst performing ASN. This also explains the 100% blacklisting of open proxies belonging to ASNs 4837 and 45758. Also, notice that all, except two ASNs, have greater than 99% blacklisting rate. Additionally, ASN 121090 has almost 32% of its open proxies involved in attacks. Moreover, more than 5% of the open proxies in ASN 7713 are vulnerable to future spam activities. On the other hand, Table 10 shows the top 10 ASNs with most blacklisted IPs around the world. Notice that, residential proxies follow trends very similar to open proxies. Particularly, all the ASNs in the table have more than 90% blacklisting rate, and eight out of ten have more than 99% blacklisted IPs. Additionally, ASN 24560, with 99.96% blacklisting, has 32.5% residential IPs proven to be involved in spamming and ≈4% of its IPs vulnerable to future spam campaigns. A common

TABLE 9. AS-level distribution of blacklisted open proxies. The number of blacklisted proxies is proportional to the total number of proxies within the AS. As shown, multiple ASs have a blacklisting rate of 100%, for instance, ASN 4134 contains 13.80% of the blacklisted open proxies. Note that most of the reported ASNs are in China and Thailand.

Rank	ASN	Blacklisted		Spam		Attack		Vulnerable	
		#	%	#	%	#	%	#	%
1	4134	114,116	100.00%	54,382	47.66%	25,052	21.95%	1211	1.06%
2	23969	46,241	99.86%	11,006	23.77%	3,624	7.83%	279	0.60%
3	7713	38,799	99.08%	6,439	16.44%	185	0.47%	1,998	5.10%
4	209	25,903	84.65%	342	1.12%	42	0.14%	20	0.07%
5	4837	25,284	100.00%	6,073	24.02%	1,642	6.49%	172	0.68%
6	3462	24,754	98.85%	4,836	19.31%	268	1.07%	65	0.26%
7	45758	19,802	100.00%	3,907	19.73%	1,037	5.24%	50	0.25%
8	8048	15,329	99.67%	1,795	11.67%	28	0.18%	63	0.41%
9	131090	14,357	99.98%	8,036	55.96%	4,548	31.67%	188	1.31%
10	28573	13,380	97.95%	843	6.17%	133	0.97%	99	0.72%
Total	Worldwide	827,106	79.11%	295,152	28.23%	72,914	6.97%	21,035	2.01%

TABLE 10. AS-level distribution of blacklisted residential proxies. The number of blacklisted proxies is proportional to the total number of proxies within the AS. Here, ASN 47331 contains 4.09% of the blacklisted residential proxies, with a blacklisting percentage of 99.04%.

Rank	ASN	Blacklisted		Spam		Attack		Vulnerable	
		#	%	#	%	#	%	#	%
1	47331	233,215	99.04%	14,901	6.33%	85	0.04%	482	0.20%
2	8151	154,078	99.84%	8,162	5.29%	17	0.01%	493	0.32%
3	9829	150,576	99.75%	50,889	33.71%	165	0.11%	8,310	5.51%
4	9121	117,330	98.42%	8,104	6.80%	49	0.04%	1,177	0.99%
5	7713	101,337	99.30%	16,496	16.16%	67	0.07%	5,931	5.81%
6	3320	97,410	99.35%	1,297	1.32%	10	0.01%	139	0.14%
7	24560	94,204	99.96%	30,618	32.49%	89	0.09%	3,731	3.96%
8	25019	80,683	99.45%	13,462	16.59%	46	0.06%	2,317	2.86%
9	12389	74,824	98.99%	11,216	14.84%	146	0.19%	1,618	2.14%
10	2856	74,274	94.31%	1,761	2.24%	30	0.04%	39	0.05%
Total	Worldwide	5,700,244	86.04%	1,081,779	16.85%	17,596	0.27%	165,328	2.58%

denominator among the residential IPs is the low proven attack record, despite the huge number of residential IPs in our dataset, and huge representation in spam activities.

Takeaways. Although the United States is the country with the third largest number of blacklisted open proxies, that only represents 55.50% of all of its open proxies, making it the country with the least percentage of blacklisted open proxies in comparison with the total number of proxies it hosts. We also observed that Indonesia and Australia have the highest number of vulnerable proxies that may lead to them being used for spamming and malicious activities in the future. Moreover, it is shown that both countries have a high percentage of proxies involved in spamming attacks; 34.59% for Indonesia, and 20.49% for Australia. Among the 99.86% of the open proxies in Nanjing that are blacklisted, more than 90% have been involved in proven spam activities, highlighting possible geographical concentration of malicious efforts. In addition, several cities and ASes have proxies blacklisting ratios of higher than 99%, indicating a possible regional blacklisting behavior.

IV. DATA ANALYSIS

A. LOCALITY CHARACTERIZATION

This study highlights the distribution of Internet proxy across countries and cities around the globe. We aim to define the relationship between such locality distribution and the characteristics of countries in terms of performance, policies, and political stability. In particular, we study the correlation between the proxy locality distribution and five characteristics, namely: censorship, Internet freedom (best and worst), political stability, Internet speed, and the country's GDP. We report the correlation using three correlation measures, namely: Pearson, Spearman, and Kendall's Tau correlation methods.

B. CORRELATION MEASURES

Correlation is a measure used to describe the relationship between two or more features in a given dataset as well as the direction of the relationship (*i.e.,* positively or negatively related). It highlights both the strength of the relationship and its direction whether it's a positive or a negative correlation. The correlation coefficient can be expressed as a value between -1 and +1. As the correlation coefficient value goes towards +1 or -1, it is an indication of either positive or negative correlation, while a correlation coefficient value around 0 means that there is no correlation between the given features. There are three types of correlations that are commonly used for measuring such relationship among independent features, namely Pearson, Spearman, and Kendall correlation.

Pearson Correlation. Pearson correlation is a correlation statistic that measures the degree of the relationship between two linearly related features using the following formula:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}. \quad (1)$$

where r_{xy} denotes Pearson r correlation coefficient between feature x and feature y, n represents the number of samples in a given dataset, x_i values of x for the i^{th} sample, and y_i represents the values of y for the i^{th} sample.

Spearman Rank Correlation. Spearman correlation is a correlation measure that is equal to the Pearson correlation between the rank values of those two features. While Pearson's correlation measures linear relationships, Spearman's correlation measures whether linear and non-linear relationships. The following formula is used to calculate the Spearman rank correlation:

$$\rho = 1 - \frac{6\sum d_i^2}{n\left(n^2-1\right)}. \quad (2)$$

where ρ means Spearman rank correlation and d_i represents the difference between the ranks of corresponding variables, n represents the number of samples.

Kendall Rank Correlation. Kendall correlation is a non-parametric test that measures the dependency strength between two features. It is used as an alternative to Pearson's correlation (parametric) when the data failed one or more assumptions of the test or when the sample size is small and has many tied ranks. The following formula is used to calculate the value of Kendall rank correlation:

$$\tau = \frac{n_c - n_d}{n(n-1)/2}. \quad (3)$$

where n_c represents the number of concordant and n_d represents the number of discordant.

C. DATA HANDLING AND PREPROCESSING

Since the range of values in the data varies widely, some of the measurements may not work properly without normalization. For example, If one of the features has a wide range of values, this may cause a failure in some of the statistical measures. Therefore, the range of all features should be normalized to be in the same range so that each feature contributes approximately proportionately to the final result.

Normalization. Data normalization is a method used to scale a set of independent values into a predefined range of values mostly from 0 to 1, without distorting differences in the ranges of values. There are many functions that can be used to perform such scaling such as min-max normalization and z-score normalization. In our measurement, we are utilizing min-max normalization method to rescale the data to be in the range [0, 1] using the following formula:

$$x_{new} = \frac{x_{old} - x_{min}}{x_{max} - x_{min}}. \quad (4)$$

where x_{old} is the original value of x and x_{new} is the normalized value of x. x_{min} and x_{max} are the maximum and minimum values in the given dataset.

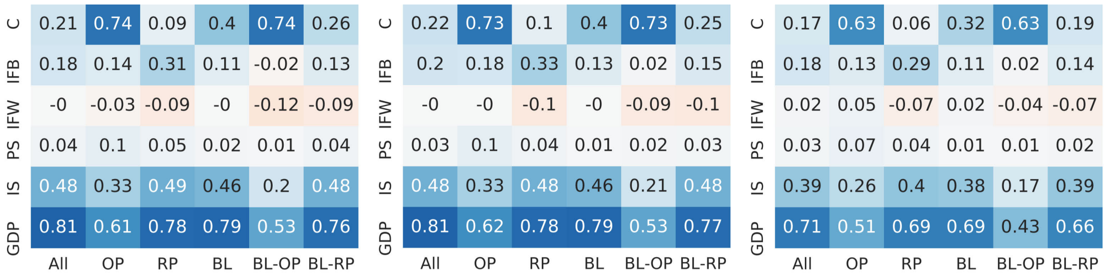
Discretization. Discretization is the process of transferring continuous values into pre-defined label interval. In this study, we mapped the continuous data to five discrete values, *i.e.,* from 0.2 to 1.0 with a distance of 0.2, representing the high end of the interval in which the data occurs, *e.g.,* values within the ranges $[0, 0.2]$ and $]0.2, 0.4]$ are assigned the values 0.2 and 0.4, respectively, and so on.

D. PROXY ANALYSIS

Censorship. Tech.co [26], a media resource for tech news and product reviews, have provided a list of the 30-most Internet-censored countries based on the monitoring policies and exposure of people to Internet contents and privacy tools (*e.g.,* VPNs). It lists Turkmenistan, North Korea, China, Eritrea, and Iran as the five most Internet-censored countries. We study the correlation between the countries policies on Internet censorship and the locality distribution of proxies. Figure 5 shows a strong positive correlation between censorship and the number of open proxies within countries. This correlation is observed for the 30-most Internet-censored countries and the locality distribution of proxies in our dataset. Since China has 16.21% of the total open proxies in the dataset, this correlation might be derived by this distribution. Generally, censorship does not show correlation with the distribution of proxies as the correlation score on Pearson measure is 0.21 for the total distribution of proxies.

Internet Freedom. According to the Freedom of the Net 2019 report [27], Iceland, Estonia, Canada, Germany, and the United States are highest with respect to Internet freedom, while China, Iran, Syria, Cuba, and Vietnam are perceived as the worst. This report is established based on a study that includes 70 analysts and 21 questions addressing the Internet access, freedom of expression, and other privacy aspects. We obtained the entire list of countries based on their ranking on Internet freedom. To demonstrate the relationship between Internet freedom and the locality distribution of proxies, we measured the correlation between the best and worst perceived 30-countries in Internet freedom and the number of proxies. Figure 5 shows that there is no correlation between Internet freedom and the distribution of proxies.

Political Stability. To explore the correlation between the distribution of proxies and countries political stability, we obtained the full list of countries ranking of political stability from the World Bank. The World Bank, the largest sources of funding and knowledge for developing countries, provides a ranking of countries based on their political stability measured by an index with values between -3 (weak) to



(a) Pearson Correlation (b) Spearman Correlation (c) Kendall Correlation

FIGURE 5. The correlation values of the number of proxies and the ranking of the country. C: censorship, IFB: Internet freedom (best 30), IFW: Internet freedom (worst 30), PS: political stability, IS: Internet speed, GDP: Gross domestic product, All: open and residential proxies, OP: open proxies, RP: residential proxies, BL: blacklisted proxies, BL-OP: blacklisted open proxies, BL-RP: blacklisted residential proxies.

2.5 (strong). The highest political stability score is assigned to Monaco (*i.e.,* 1.61 points), while the lowest score is assigned to Yemen (*i.e.,*-3 points). We observe that the score of -3 is given to Yemen by the source of data as a sign of data unavailability or severe political instability due to an ongoing war in the region. The results in Figure 5 show that there is no observed correlation between countries political stability and the distribution of proxies.

Internet Speed. We explore the correlation between Internet speed and distribution of proxies. We obtained the list of countries ranking of Internet speed from Speedtest by Ookla [28]. We observed a positive correlation between Internet speed and the distribution of proxies in general. This positive correlation is also observed with the blacklisted residential proxies. This is inline with the intuitive that one can expect the distribution of proxies locality fits positively with the Internet speed.

GDP. We finally explore the correlation between the GDP and the distribution of proxies. The GDP is a monetary indicator that measures of the market value based on the production of all goods and services in a certain time period. We obtained the countries GDP ranking data from the International Monetary Fund [29] for the year 2020. The analysis shows a strong positive correlation between the GDP and proxies localities, especially for the residential proxies. This is due to the fact that countries with higher GDP often maintain high operational services to host Internet proxies.

V. CONCLUSION

Internet proxies are intermediary and a gateway between users and servers, often used to protect users' privacy and hide their identity. Moreover, proxies are used to surpass the policies-enforced regional restrictions on accessing the Internet, enabling the user's freedom use of the Internet. However, they may be used by adversaries to launch attacks, collect users' data, and inject ads and files. In this study, we highlight this by conducting a comprehensive study on two types of proxies, *i.e.,* open and residential proxies. By studying a dataset of 1,045,468 open proxies and 6,419,987 residential proxies, we found that 79.11% of the open proxies are blacklisted via different blacklisting services, with 28.23% labeled as spam proxies, and 6.97% labeled as proxies used to launch an attack. Similarly, our analysis shows that 86.04% of the residential proxies are blacklisted, despite their efforts in hiding their identity, with 16.85% labeled as spam and 0.27% are associated with an adversary attacks. Further, we found that the distribution of the proxies is positively correlated with the GDP and Internet speed on the country-level of residence. While Internet proxies are considered a privacy preserving way to access the Internet, this study, along with several studies in the literature, highlights the malicious use of the proxies, and the risk of using them.

REFERENCES

[1] M. Mukherjee, R. Matam, L. Shu, L. Maglaras, M. A. Ferrag, N. Choudhury, and V. Kumar, "Security and privacy in fog computing: Challenges," *IEEE Access*, vol. 5, pp. 19293–19304, 2017.

[2] S. Yu, "Big privacy: Challenges and opportunities of privacy study in the age of big data," *IEEE Access*, vol. 4, pp. 2751–2763, 2016.

[3] Avast. (2020). *What is a Proxy Server and How Does it Work?* Accessed: Nov. 2020. [Online] Available: https://www.avast.com/c-what-is-a-proxy-server

[4] J. Castellà-Roca, A. Viejo, and J. Herrera-Joancomartí, "Preserving user's privacy in Web search engines," *Comput. Commun.*, vol. 32, nos. 13–14, pp. 1541–1551, Aug. 2009.

[5] H. Yu, E. Lee, and S.-B. Lee, "SymBiosis: Anti-censorship and anonymous Web-browsing ecosystem," *IEEE Access*, vol. 4, pp. 3547–3556, 2016.

[6] Didsoft. (2019). *Free Proxy List*. Accessed: Nov. 2019. [Online] Available: https://bit.ly/2vzLCYI

[7] ProxyNova.com. (2019). *Nova Proxy Switcher*. Accessed: Nov. 2019. [Online] Available: https://bit.ly/2OguC0d

[8] X. Mi, X. Feng, X. Liao, B. Liu, X. Wang, F. Qian, Z. Li, S. Alrwais, L. Sun, and Y. Liu, "Resident evil: Understanding residential IP proxy as a dark service," in *Proc. IEEE Symp. Secur. Privacy (SP)*, San Francisco, CA, USA, May 2019, pp. 1185–1201.

[9] W. Scott, R. Bhoraskar, and A. Krishnamurthy, "Understanding open proxies in the wild," in *Proc. Chaos Commun. Camp*, Aug. 2015, pp. 1–11.

[10] K. Steding-Jessen, N. L. Vijaykumar, and A. Montes, "Using low-interaction honeypots to study the abuse of open proxies to send Spam," *INFOCOMP J. Comput. Sci.*, vol. 7, no. 1, pp. 44–52, 2008.

[11] G. Tyson, S. Huang, F. Cuadrado, I. Castro, V. C. Perta, A. Sathiaseelan, and S. Uhlig, "Exploring HTTP header manipulation in-the-wild," in *Proc. 26th Int. Conf. World Wide Web*, Perth, WA, Australia, Apr. 2017, pp. 451–458.

[12] S. Kanchan and N. S. Chaudhari, "SRCPR: SignReCrypting proxy re-signature in secure VANET groups," *IEEE Access*, vol. 6, pp. 59282–59295, 2018.

[13] A. Mani, T. Vaidya, D. Dworken, and M. Sherr, "An extensive evaluation of the Internet's open proxies," in *Proc. 34th Annu. Comput. Secur. Appl. Conf.*, Dec. 2018, pp. 252–265.

[14] G. Tsirantonakis, P. Ilia, S. Ioannidis, E. Athanasopoulos, and M. Polychronakis, "A large-scale analysis of content modification by open HTTP proxies," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, San Diego, CA, USA, 2018, pp. 1–15.

[15] D. Perino, M. Varvello, and C. Soriente, "ProxyTorrent: Untangling the free HTTP(S) proxy ecosystem," in *Proc. World Wide Web Conf. World Wide Web (WWW)*, Lyon, France, Apr. 2018, pp. 197–206.

[16] T. Chung, D. Choffnes, and A. Mislove, "Tunneling for transparency: A large-scale analysis of End-to-End violations in the Internet," in *Proc. ACM Internet Meas. Conf. (IMC)*, Santa Monica, CA, USA, Nov. 2016, pp. 199–213.

[17] Z. Weinberg, S. Cho, N. Christin, V. Sekar, and P. Gill, "How to catch when proxies lie: Verifying the physical locations of network proxies with active geolocation," in *Proc. Internet Meas. Conf.*, Boston, MA, USA, Oct./Nov. 2018, pp. 203–217.

[18] N. Weaver, C. Kreibich, M. Dam, and V. Paxson, "Here be Web proxies," in *Proc. 15th Int. Conf. Passive Act. Meas. (PAM)*, Los Angeles, CA, USA, Mar. 2014, pp. 183–192.

[19] IP2Location.com. (2019). *IP2Proxy*. Accessed: Nov. 2019. [Online] Available: https://bit.ly/2RJGgmg

[20] M Developers. (2019). *MultiProxy*. Accessed: Nov. 2019. [Online] Available: https://bit.ly/2UdNoJp

[21] P List Developers. (2019). *Proxy List*. Accessed: Nov. 2019. [Online] Available: https://bit.ly/2OePuEU

[22] Checkerproxy.Net. *Checkerproxy*. Accessed: Nov. 2019. [Online] Available: https://bit.ly/31cMnTn

[23] P Developers. (2019). *Proxybroker*. Accessed: Nov. 2019. [Online] Available: https://bit.ly/2S54ir6

[24] MaxMind. (2019). *Maxmind*. Accessed: Nov. 2019. [Online]. Available: https://www.maxmind.com/

[25] M. Wander, C. Boelmann, L. Schwittmann, and T. Weis, "Measurement of globally visible DNS injection," *IEEE Access*, vol. 2, pp. 526–536, 2014.

[26] (2019). *Tech.Co: Internet Censorship Rankings*. Accessed: Feb. 2020. [Online] Available: https://tech.co/vpn/internet-censorship-rankings

[27] (2019). *Freedom on The Net: Global Internet Freedom Ranking*. Accessed: Feb. 2020. [Online] Available: https://www.freedomonthenet.org/report/freedom-on-the-net/2019/the-crisis-of-social-media

[28] (2020). *Ookla: Internet Speeds By Country*. Accessed: Feb. 2020. [Online] Available: https://www.speedtest.net/global-index

[29] (2020). *International Monetary Fund: Countries GDP Ranking*. Accessed: Feb. 2020. [Online] Available: https://www.imf.org/

JINCHUN CHOI received the B.E. and M.S. degrees from Inha University, in 2011 and 2014, respectively. He is currently pursuing the Ph.D. degree with the Department of Computer Science, University of Central Florida and the Department of Computer Information Science, Inha University (joint Ph.D. program). His research interests include networks and the IoT security.

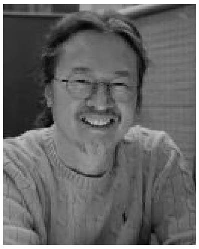
MOHAMMED ABUHAMAD received the B.S. degree in computer science from The IUG, in 2007, and the M.S. degree in artificial intelligence from The National University of Malaysia, in 2013. He is currently pursuing the Ph.D. degree with the Information Security Research Laboratory (ISRL), Inha University, South Korea, and the Security Analytics Research Lab (SEAL), University of Central Florida. His research interests include software security, machine learning, authentication, privacy, and deep learning.

AHMED ABUSNAINA (Graduate Student Member, IEEE) received the B.Sc. degree in computer engineering from An-Najah National University, Palestine, in 2018. He is currently pursuing the Ph.D. degree with the Department of Computer Science, University of Central Florida. His research interests include software security, machine learning, and adversarial machine learning.

AFSAH ANWAR (Graduate Student Member, IEEE) received the B.S. degree from Jamia Millia Islamia University, New Delhi, India, in 2014. He is currently pursuing the Ph.D. degree with the Department of Computer Science, University of Central Florida. Before starting his Ph.D. degree, he was working as a Data Analyst (C) for Apple. His research interests include binary analysis, vulnerability analysis, and malware analysis.

SULTAN ALSHAMRANI received the B.S. degree in computer science from the University of Tabuk, Saudi Arabia, in 2014, and the M.S. degree in computer science from Loyola University Chicago, Illinois, USA, in 2018. He is currently pursuing the Ph.D. degree with the Security Analytics Research Lab (SEAL), University of Central Florida. His research interests include data mining, natural language processing, and deep learning.

JEMAN PARK received the B.Sc. degree in computer and communication engineering from Korea University, Seoul, South Korea, in 2016. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, University of Central Florida. His work has been focused on privacy, computer security, and systems.



DAEHUN NYANG received the B.Eng. degree in electronic engineering from the Korea Advanced Institute of Science and Technology, and the M.S. and Ph.D. degrees in computer science from Yonsei University, South Korea, in 1994, 1996, and 2000, respectively. He was a Senior Member of the engineering staff at the Electronics and Telecommunications Research Institute, South Korea, from 2000 to 2003. Since 2003, he has been a Full Professor with the Computer Information Engineering Department, Inha University, South Korea, where he is also the founding Director of the Information Security Research Laboratory. His research interests include AI-based security, network security, traffic measurement, privacy, usable security, biometrics, and cryptography. He is a member of the board of directors and an editorial board of *ETRI Journal* and also Korean Institute of Information Security and Cryptology.

DAVID MOHAISEN (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees from the University of Minnesota, in 2012. He is currently an Associate Professor with the University of Central Florida, where he directs the Security and Analytics Lab (SEAL). Before joining UCF, in 2017, he was an Assistant Professor with SUNY Buffalo, from 2015 to 2017, and a Senior Research Scientist with the Verisign Labs, from 2012 to 2015. His research interests are in the areas of networked systems and their security, online privacy, and measurements. He is a Senior Member of ACM (2018). He is an Editor-in-Chief of *EAI Transactions on Security and Safety*, and an Associate Editor of the IEEE TRANSACTIONS ON MOBILE COMPUTING, *Elsevier Computer Networks*, and *ETRI Journal* (Wiley).