WhoisXMLAPI
The Who Behind Domain, IP & Cyber Threat Intelligence

WHITE PAPER

The Secret Life of Wild Subdomains

Looking into the subdomain surface of today's most spoofed brands

Contents

1

Executive Summary

Managing attack surfaces is a critical cybersecurity process, though it is not always an easy one. Some areas in an organization's attack surface could remain undocumented, unnoticed, and unseen. Certain subdomains are among these potentially hidden threat sources that might be forgotten, hijacked, and maliciously used as a result.

In this study, we gathered data on and analyzed the subdomain surface of 10 of today's most-spoofed brands—PayPal, Facebook, Microsoft, Netflix, WhatsApp, Bank of America, CIBC, Desjardins, Apple, and Amazon.

Here are our high-level findings:

- Almost half of the associated root domains are less than two years old according to their WHOIS records, while thousands were recently registered.
- "Blog," "host," and "master" are among the most frequently used terms in the subdomains, aside from the brand names.
- Some 56% of the subdomains in the sample data belong to the .com top-level domain (TLD) space.
- The U.S. is the top registrant country (42%), followed by China (14%), Canada (10%), and Panama (7%), according to the domains' WHOIS records.
- The top locations of the IP addresses linked to the subdomains are the U.S. (54%), Germany (6%), Ireland (5%), and Canada (4%).
- Publicly attributable domains only make up 0.17% of the total number of root domains in the sample data.
- Only 8.3% of the subdomains appear on PhishTank, leaving a considerable percentage undocumented.
- Typosquatting data feeds can amplify the number of root domains identified in registration groups by as much as 146 times.
- Some 79% of the domain names use the brand's name in third-level domains, while the rest use them in the root domains.

2 Key Concepts Clarified

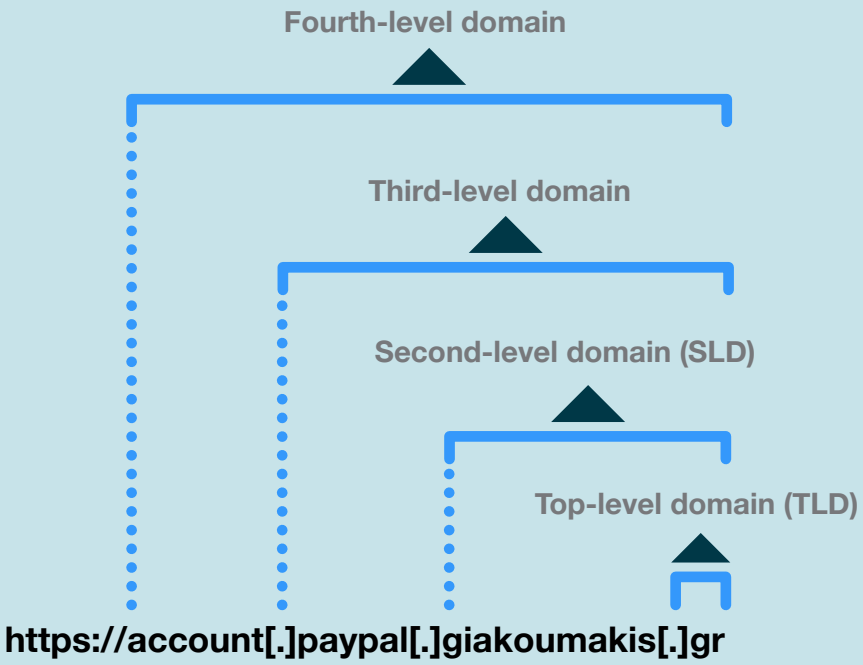
This study delves into several concepts, including domain attack surfaces, root domains, subdomains, and connected domain footprints. Each concept is briefly described below.

2.1 Domain Attack Surface

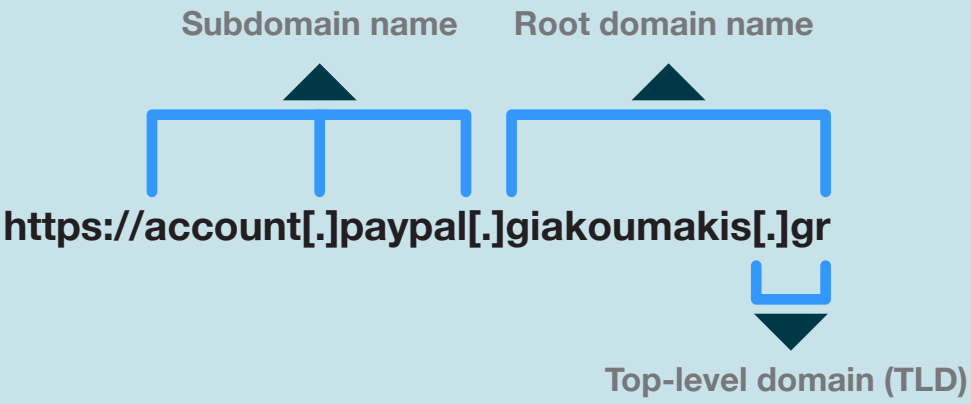
An organization's domain attack surface refers to the set of all domains and subdomains that could cause it harm in some way, particularly if they are related to its brand name. Such domains and subdomains can possibly figure in malicious activities, such as cybersquatting, phishing, and malware campaigns, thereby contributing to the vulnerability of the spoofed brand to attacks.

2.2 Subdomains versus Root Domains

For the purpose of this study, we consider root domains as second-level domains (e.g.,giakoumakis[.]gr) and subdomains as third - and fourth-level domains (e.g., account[.]paypal[.]giakoumakis[.]gr), as shown in the illustration below. For completeness, it is worth noting that some second-level domains like ".co.uk" act as top-level ones and are dealt properly within our sample.



For the sake of simplicity, we will, however, most commonly refer to subdomains and root domains as follows.



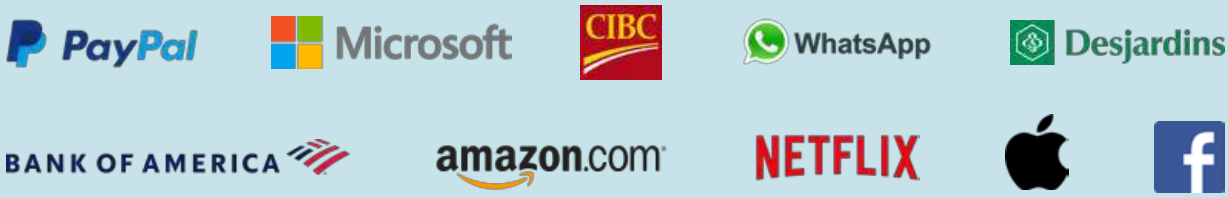
Many cybersecurity research methods addressing threats like brand spoofing or typosquatting focus on the root domain name, just like the domain name business. In this study, however, we shall focus on the subdomain names and illustrate their important share in domain surfaces.

2.3 Domain and Subdomain Footprints

Domain and subdomain footprints refer to all domains and subdomains with one or more elements recorded in common. Elements in this study will include text strings (e.g., a brand name or close variation of it) and shared IP address resolutions obtained from passive Domain Name System (pDNS) records. The latter means that two or more different domain names have been observed to resolve to the same IP address.

3 Data Collection and Tools

This study intends to analyze and investigate the hidden domain attack surfaces of 10 of the most frequently spoofed brands, namely:



A combination of WHOIS, DNS, and IP intelligence tools and databases was used to gather and analyze each brand's domain attack surface. Specifically, a pDNS database containing 20GB of highly compressed data was downloaded, where the covered data points include domain names, IP addresses, and timestamps. The domain names typically included a subdomain name, albeit some root domains may also appear directly.

Overall, each point describes an event at the given time when the domain was observed to resolve to a given IP address in the DNS by a pDNS sensor. In our data set, most of these events occurred in September 2020. From the said data, more data sets were derived, including:

- Records where the brand names appeared in any part of the domain name
- Root domains of the domains in the sample data set
- Most commonly used words that appeared alongside the domain names in the sample

Additional tools and databases were used to process the domain names and IP addresses of the sample, including:

- **IP Geolocation API:** To derive the geographic location of potential threat sources' infrastructure based on the IP addresses shown in pDNS records.
- **WHOIS Database Download:** To determine the root domains' registrant country and age.
- **Typosquatting Data Feed:** To see the presence of the sample data's root domains in typosquatting groups.
- **PhishTank:** To know if the domain names in our data set are part of PhishTank's list of confirmed phishing URLs.

4 Findings

4.1 Average Size of the Brands' Domain Attack Surface

Based on the sample, the average size of the 10 brands' domain attack surface comprises 17,734 domains and subdomains. The table below shows each brand's domain attack surface size led by Apple with 54,187 connected web properties. Facebook has the second largest domain attack surface with 32,547 connected properties, closely followed by Amazon with 31,570.

Brand	Domain Attack Surface Size
Apple	54,187
Facebook	32,547
Amazon	31,570
Microsoft	26,885
PayPal	11,075
WhatsApp	9,278
Netflix	4,631
Bank of America	4,018
Desjardins	2,075
CIBC	1,076
Total number of domains and subdomains	**177,342**
Average	**17,734**

4.2 Age of Domains

Considering root domains and their WHOIS data, a total of 29,171 had registration dates between January 2011 and October 2020. Almost half (45%) of these are barely two years old, while 5,074 are newly registered. The table below shows the total number of root domains containing the subdomains for each brand from 2011 to 2020.

	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Apple	646	611	631	541	589	773	791	886	4,596	1,763
Facebook	539	366	360	387	419	476	459	449	715	801
Amazon	288	260	263	308	399	417	481	539	1256	1,226
Microsoft	71	88	66	73	86	102	93	116	162	149
PayPal	105	119	123	140	143	150	134	191	420	535
WhatsApp	86	98	130	147	186	229	301	364	556	317
Netflix	49	50	45	52	49	70	100	101	245	232
Bank of America	19	16	17	14	33	27	36	20	40	28
Desjardins	6	7	8	5	10	4	10	11	8	4
CIBC	13	9	10	10	13	8	10	12	36	19
Total	**1,822**	**1,624**	**1,653**	**1,677**	**1,927**	**2,256**	**2,415**	**2,689**	**8,034**	**5,074**

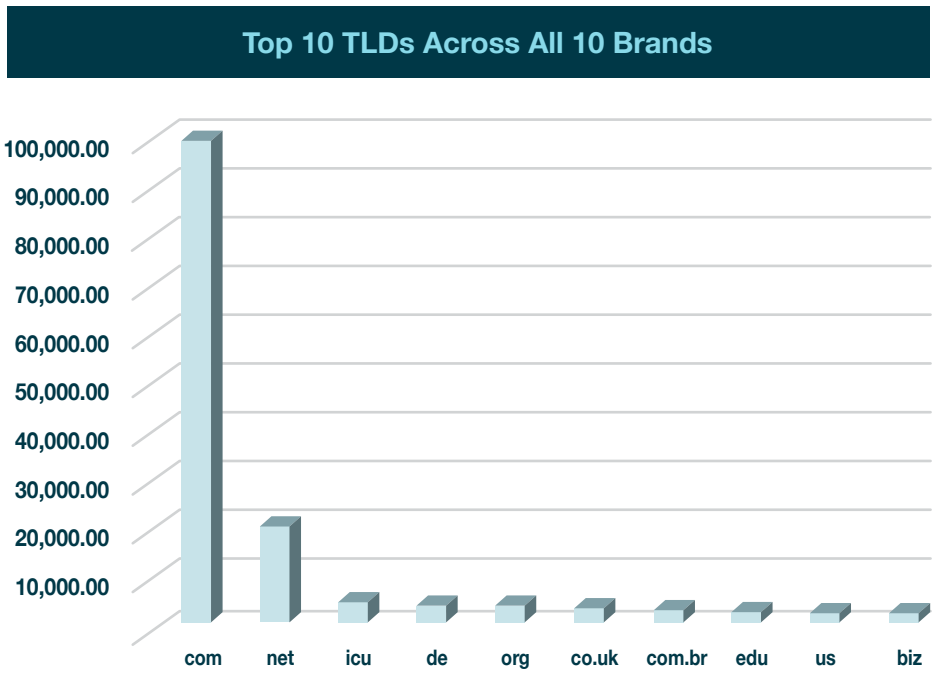
4.3 Commonly Appearing Terms

These are the top 10 terms that frequently appeared alongside the brand names in the full domain names (including in the root domains and/or subdomains)

Blog | Host | Master | Office | Service
Online | Mail | Account | Bank | Edge

4.4 Top TLDs Used for the Root Domains and Subdomains

The most commonly used TLD for root domains is .com, numbering 99,242—equivalent to 56% of the total sample size. The rest of the domains are distributed across 497 other TLDs. The chart below shows the top 10 TLDs used for all 10 brands.

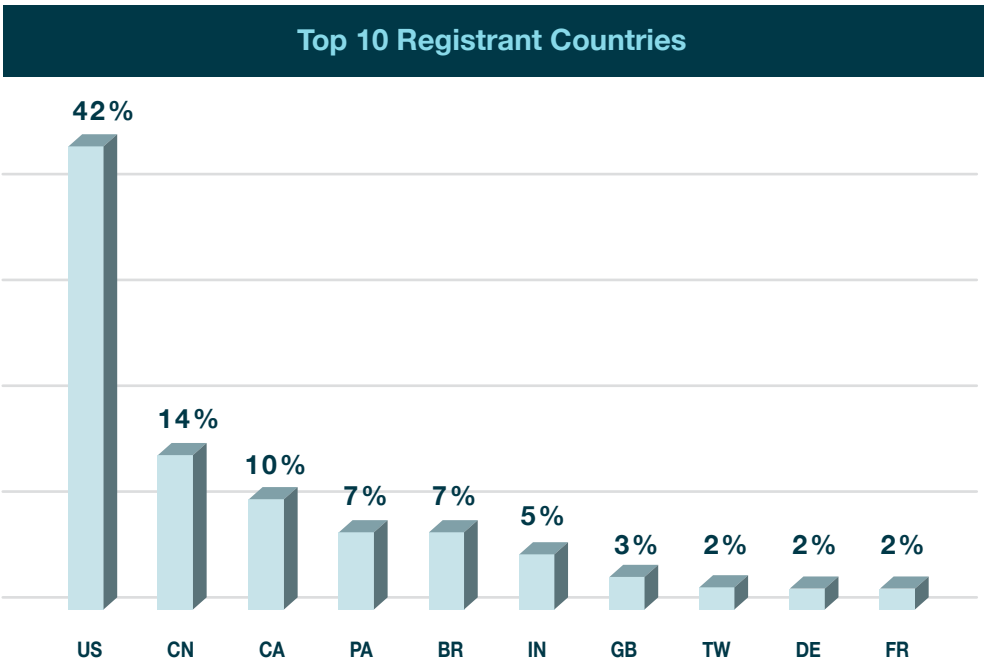


A similar trend can be observed for each brand, especially for the .com TLD. See the table below that shows the top 3 TLDs for each brand

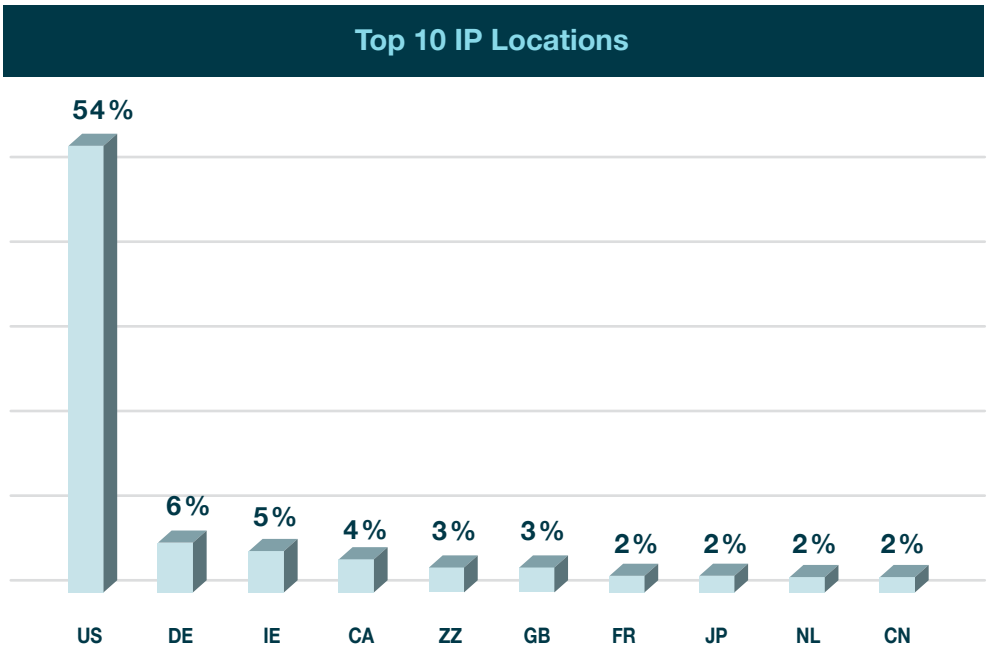
Brand	Top 1 TLD	Top 2 TLD	Top 3 TLD
Amazon	.com	.net	.edu
Apple	.com	.icu	.net
Bank of America	.com	.net	.br
CIBC	.com	.pl	.ca
Desjardins	.com	.ca	.ws
Facebook	.com	.net	.de
Microsoft	.com	.net	.org
Netflix	.com	.net	.live
PayPal	.com	.net	.biz
WhatsApp	.com	.net	.br

4.5 Domain and Subdomain Location

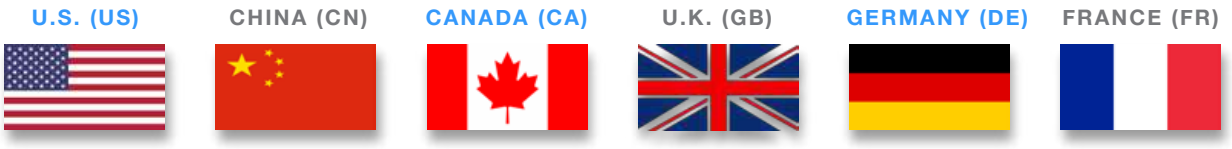
The full WHOIS records of the root domains reveal that 42% cited the U.S. as their registrant country. China took the second spot, used by 14% of the subdomains. The chart below shows the subdomain distribution across the top 10 registrant countries.



IP geolocation findings, which refer to IP data and infrastructure location, reveal that most subdomains (54%) can be traced back to the U.S. In part, this might be due to transient instances where certain registrars offer a free domain name with privacy protection included to attract new registrants. The rest of the top 10 registrant countries based on IP geolocation account for 6% or less of the subdomains. Refer to the chart below for details.



Six countries can be consistently seen in both the top 10 registrant countries according to WHOIS and IP geolocation data, namely:

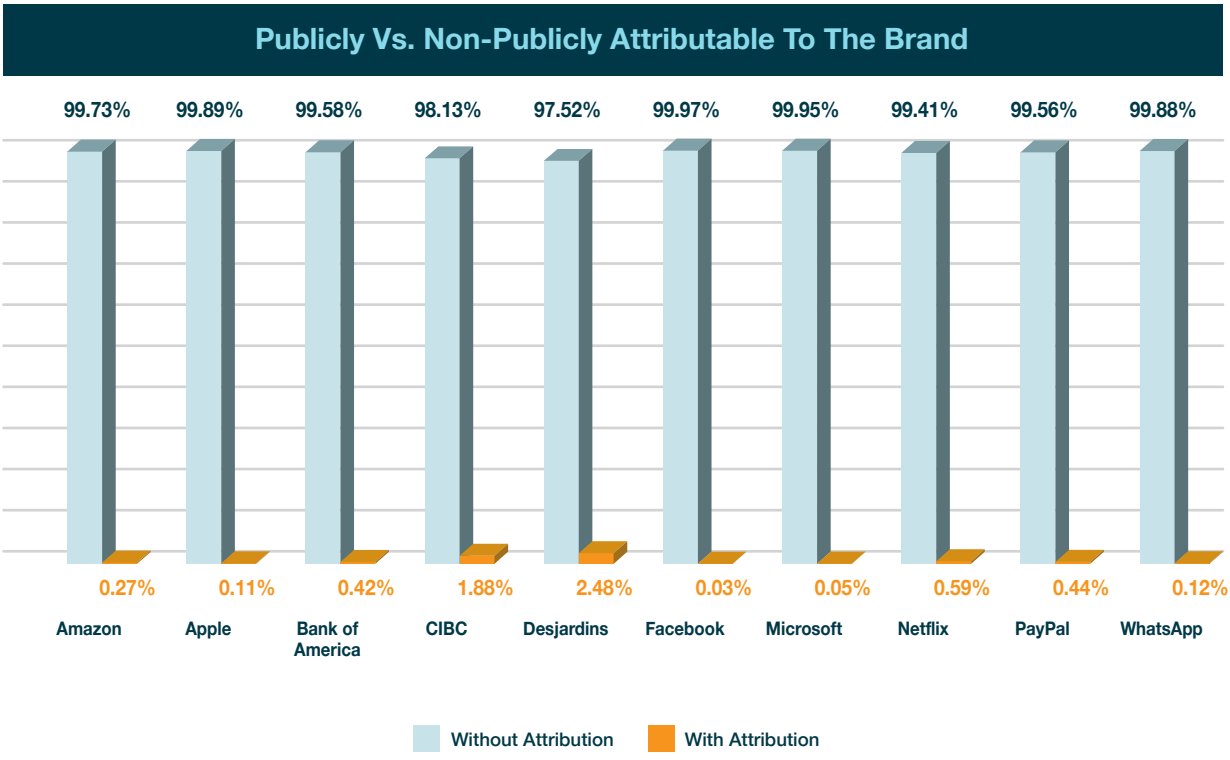


4.6 Publicly versus Non-Publicly Attributable Domains

This study distinguishes between attributable and non-attributable second-level domains based on whether the root domains derived from the sample data share the same registrant email address as the one present in the WHOIS records of the brand's official root domain.

The latest WHOIS record of amazon[.]com, for example, contains the registrant email address hostmaster@amazon[.]com. Any domain or subdomain with the string "amazon" is considered publicly attributable if its current WHOIS record includes the same registrant email address. Any domain or subdomain without this "proof" is considered non-attributable in our context.

As such, the percentage of domains proven attributable is very low for each brand, ranging between 0.03% and 2.48%. On average, non-attributable domains make up more than 99.8% of the brands' total surface.

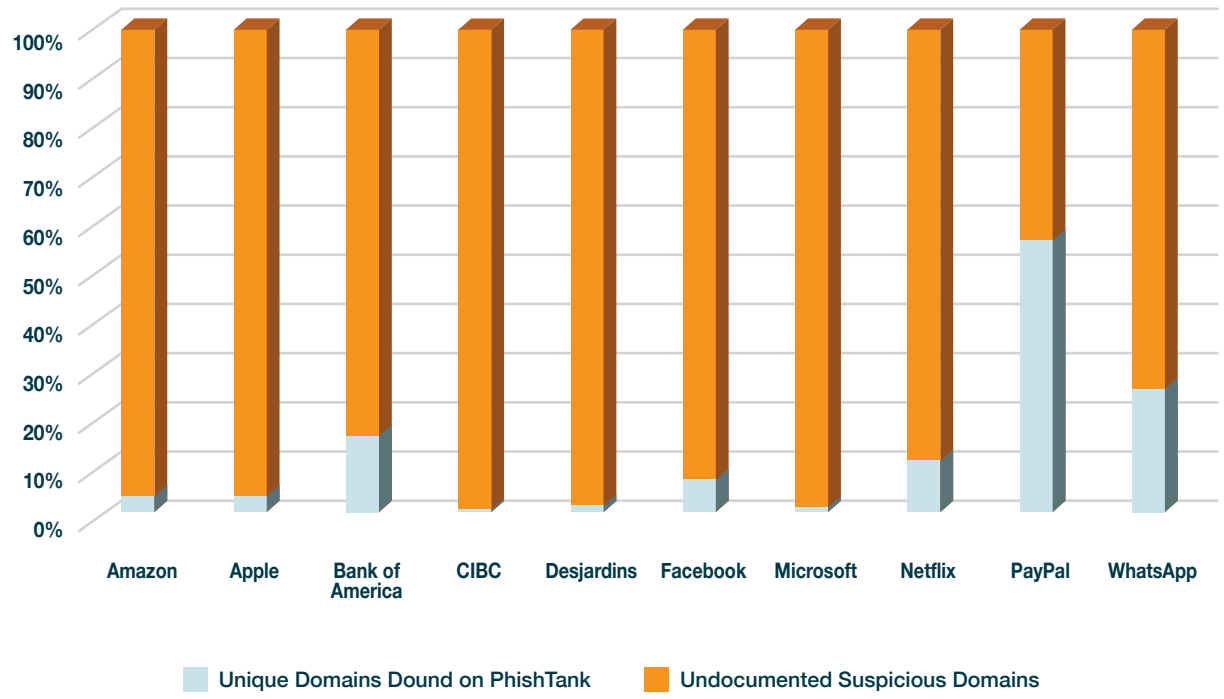


4.7 Phishing Domains Detected

As part of the study's next step, a list of verified phishing URLs dated between 1 June and 16 October 2020 was downloaded from PhishTank. These URLs were compared against the 10 brands' domain attack surface. Around 8.3% of the subdomains in this study was found on PhishTank since the corresponding domain name was part of a phishing URL.

The following chart shows the ratio between the domains found on PhishTank and the undocumented domains that may comprise each brand's unknown attack surface.

Ratio of Unique Domains Found on PhishTank vs. Undocumented Suspicious Domains



At 55.9%, PayPal has the highest number of subdomains found on PhishTank. However, almost all brands have a large percentage of undocumented domains. On average, less than 1% of the domains are listed on PhishTank though they share similar characteristics.

4.8 Typosquatting Domains Detected

From our sample, a total of 369 unique root domains were identified by the typosquatting data feeds, dated 1 October 2019 to 9 October 2020. The feed lists groups of root domain names that were registered on the same day and whose name is similar to others within the same group. These domains belong to different typosquatting groups, totaling 29,621 look-alike domain names.

The table below shows the breakdown for each brand along with the total number of unique domains that could be found as members of the typosquatting data feeds' groups (except for Desjardins whose name did not appear in any group). The percentages show by how much the feeds can help amplify the number of suspicious root domains.

Brand	Number of Root Domains Present in Typosquatting Data Feeds	Number of Unique Typosquatting Domains Found	Percentage of Suspicious Domains
Amazon	196	28,722	14654.08%
WhatsApp	5	46	920.00%
Apple	105	613	583.80%
PayPal	24	125	520.83%
Netflix	8	30	375.00%
Microsoft	12	43	358.33%
CIBC	1	3	300.00%
Facebook	16	35	218.75%
Bank of America	2	4	200.00%

Presence of Brand Names in Root Domains and Subdomains

Around 21% of the domains in this study contained the brand names in the root domain, while 79% did so in the third- or fourth-level domains (i.e., subdomains). The table below shows the percentages per brand.

Brand Name	Total Number of Domains/ Subdomains	Number of Domains with the Brand Name in the Root Domain	Percentage of Domains with the Brand Name in the Root Domain	Percentage of Domains without the Brand Name in the Root Domain
WhatsApp	9,278	225	2.40%	97.60%
PayPal	11,075	1,096	9.90%	90.10%
Facebook	32,547	5,521	17.00%	83.00%
Amazon	31,570	5,522	17.50%	82.50%
Apple	54,187	10,583	19.50%	80.50%
Netflix	4,631	958	20.70%	79.30%
Microsoft	26,885	8,400	31.20%	68.80%
CIBC	1,076	469	43.60%	56.40%
Bank of America	4,018	2,791	69.50%	30.50%
Desjardins	2,075	1,740	83.90%	16.10%

5

Final Thoughts

The main purpose of this study was to research the extent to which subdomains are part of large organizations' domain attack surfaces that may arguably serve as entry points in cyber attacks. Using a combination of pDNS, WHOIS, IP, and other intelligence sources, we uncovered and analyzed a sample of 177,342 subdomains that contained the names of some of today's most-spoofed brands.

Most importantly, we found that a total of 99.83% of the domains and subdomains containing the 10 brand names couldn't be attributed to the companies.

While there are legitimate reasons for using other entities' brand names, such as running an online shop or blog about them, we found over 14,700 subdomains from the sample data on PhishTank's list of verified phishing domains. The subdomains documented on Phishtank only represents 8.3% of the total number of domains in the sample data, leaving 91.7% undocumented and possibly dangerous. Some of these subdomains were unreasonably long and made up of many layers—e.g., account[.]paypal.com[.]paypal.com[.]paypal[.]com[.][.]paypal.com[.]paypal.com[.]paypal[.]com[.]paypal[.]com[.]giakoumaskis[.]gr—and could be easily exploitable as users would hardly notice, depending on what communication mechanism is being used, that the actual root domain (i.e., giakoumaskis[.]gr) is not paypal[.]com.

We also found a total of 369 root domains appearing in both our sample and WhoisXML API's typosquatting data feed files dated 1 October 2019 – 9 October 2020. From these 369 root domains, a total of 29,621 root domains — or 81 times more—were identified in the feed files for pertaining to the same registration groups.

Understanding and managing domain attack surfaces is crucial for the cybersecurity of enterprises and end-users alike. Using a combination of databases and tools can help uncover and monitor as much as those surfaces as possible. To learn more, visit https://www.whoisxmlapi.com/.

About Us

WhoisXML API is a cyber intelligence provider that gives enterprises access to one of the largest repositories of well-parsed domain, subdomain, IP, and DNS data that enhances cybersecurity platforms' capabilities and helps security teams gain superior network security.

The data that WhoisXML API provides comes in different consumption models, ranging from APIs, data feeds, monitoring tools, and lookup tools, all of which make the Internet more secure and transparent. WhoisXML API has more than 50,000 satisfied customers, spanning law enforcement agencies, cyber forensics analysts, threat hunters, and cybersecurity solutions developers.