Feasibility of Large-Scale Vulnerability Notifications after GDPR

Wissem Soussi, Maciej Korczyński, Sourena Maroofi, Andrzej Duda
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble, France

***Abstract*—In this paper, we consider the problem of effective notifications of domain abuse or vulnerabilities to the domain owners, administrators, or webmasters. We have developed a scanner to test whether selected email aliases specified in RFC 2142 are correctly configured and whether notifications can be successfully delivered. We also test the reachability of email addresses collected from the DNS Start of Authority (SOA) records. Based on a measurement campaign of a large number of domains compared to the previous studies (4,602,907 domains), we show that domains are more reachable through SOA contacts. We find that the country-code TLD names are more reachable compared to the new gTLD names. We have also observed that the most used generic email alias is *abuse* (67.95%). Using regression analysis, we show the relationship between the reachability of email addresses and the fact that names are hosted on large shared platforms or have a significant value. Our results confirm that direct notification channels are currently not scalable, so we propose a scheme that preserves user privacy in compliance with GDPR and supports large-scale vulnerability notifications.**

1. Introduction

Malicious actors compromise thousands of legitimate domains every day by exploiting vulnerable content management systems (CMSs), frameworks, or libraries used to build websites. The compromised domains are abused to launch Internet-scale phishing, malware drive-by-download, or spam campaigns. To prevent vulnerable resources from being exploited and to remediate already abused domains, defenders share information about security threats and incidents through collaborative clearing houses such as Anti-Phishing Working Group (APWG)[1], PhishTank[2], or URLhaus (Malware URL exchange)[3].

Some types of domain abuse or vulnerabilities should be directly reported to the domain registrants (owners), administrators, or webmasters. Therefore, the Internet community needs to maintain the ability of large-scale notification mechanisms. An alternative approach is to report abuse through intermediaries such as Computer Emergency Response Teams (CERTs). They validate email notifications and further communicate with the actors responsible for the affected systems. However, previous work showed that security notifications directly addressed to the owners of vulnerable resources promote faster remediation than those sent to national CERTs [1]. Therefore, several researchers used the contacts of domain administrators and owners retrieved from the public WHOIS data [1]–[8] and studied different communications strategies to increase remediation rates.

Retrieving contact information at scale from public WHOIS is highly problematic [2] and became even more difficult with the introduction of the General Data Protection Regulation (GDPR) on May 25, 2018. The Internet Corporation for Assigned Names and Numbers (ICANN) adopted the temporary specification for generic top-level domains (gTLDs) on how to publish the registration data of individuals [9] that prohibits domain registrars and registries from storing personal data in the public WHOIS database, in particular, the email addresses of registrants and administrators. In the absence of direct contact with the registrant, it is recommended to contact the relevant registrar who has to provide access to registrant contact information in "a reasonable time" [10]. However, this rule may cause significant delays in fixing vulnerabilities and mitigating abuse, and in addition, it does not scale.

Previous studies have investigated the effectiveness and feasibility of security notifications (content verbosity, external redirection links, message language, etc.) [1]–[8]. In this paper, motivated by the implications of the EU regulation on data protection, we systematically test available direct contacts of domain owners and administrators as defined in RFC 2142 [11]. For a given domain `example.com`, RFC 2142 requires to configure valid email aliases such as `abuse@example.com` or `security@example.com` for incident and vulnerability notifications. Rather than quantifying remediation rates, we test whether five RFC-specific generic email aliases are correctly configured and whether notifications can be successfully delivered. We also test the reachability of email addresses collected from the DNS Start of Authority (SOA) resource records (RR).

We perform our measurement study on a large number of domains compared to the previous studies (4,602,907 domains). A big part of the domains are representative samples of *.com* and *.net* legacy gTLD names, country-code TLD (ccTLD) names, and new gTLD names. We also perform measurements on selected domains: i) found in Tranco top 1 M ranking list [12] of February 2020, ii) those with vulnerable WordPress plugins or versions, iii) the domains vulnerable to DNS AXFR transfers [13], iv) DNS zone poisoning [14], and v) compromised domains.

Our study shows that in all the examined categories, domains (their owners and administrators) are more reachable through SOA contacts. As expected, the results for the Tranco top 1 M list show that better-ranked domains comply with the RFC 2142 specification and are more reachable. The ccTLD names are more reachable for both SOA (35.4%) and RFC-specific aliases (12.68%) compared to the new gTLD names (21.35% and only 3.97%, respectively). For instance, 61.63% of the sampled

[1]https://apwg.org
[2]https://www.phishtank.com
[3]https://urlhaus.abuse.ch

DOI 10.1109/EuroS&PW51379.2020.00077

domains for new gTLD names have no MX configured record. Moreover, based on the email addresses stored in DNS SOA records, the domains with known vulnerable versions of WordPress are more reachable than Tranco top 1 M domains (43.38% versus 39.74%). We conclude that the domains using CMSs are often hosted on a shared facility, and therefore, it is the provider that configures the DNS SOA records. We have finally observed that the most used generic email alias is *abuse* (67.95%).

To statistically verify the different interpretations of the results, we build a Generalized Linear Model (GLM). We collect variables assuming they represent different characteristics of the domain such as the effort of the domain owner and the DNS administrator, the value of the domain, and the fact that it is hosted in a shared host or not. The models confirm that valuable websites hosted on large shared platforms are more likely to have correctly configured email addresses.

Our results confirm that direct channel notifications are currently not scalable, so we propose a scheme that preserves user privacy in compliance with GDPR and supports large-scale vulnerability notifications.

2. Methodology

We have developed a scanner to systematically test available direct contacts of domain owners and administrators. We first scan for the DNS MX records of the domain and select a mail server with the highest priority. Afterwards, we establish different connections using the Simple Mail Transfer Protocol (SMTP) [15] to the selected mail server. We do not send emails, but we verify the existence of an email address using the RCPT TO query followed by the destination email address. If the mail server replies with code 250, the recipient address is considered as valid. In a single SMTP session, we only test one email address to avoid triggering mechanisms preventing email address enumeration [15], which may close the SMTP connection and blacklist the IP address of the sender.

Another countermeasure used by mail servers is to accept any given recipient, even a non-existent one, and return code 250. This procedure is called CATCH ALL (or wildcard email address) [16]. Our scanner is designed to detect if a mail server uses the CATCH ALL mechanism by checking for the existence of a randomly generated email. If such a contact is validated, then the mail server is most likely validating all, even non-existent, addresses.

For each sampled domain name, we generate email aliases using the names defined in RFC 2142 [11]: for the domain `example.com`, we test the validity of the following email aliases: `hostmaster@example.com`, `webmaster@example.com` (for DNS and HTTP issues), `abuse@example.com` (for generic abuse and vulnerability notifications), `noc@example.com`, and `security@example.com` (for network security).

We scan for DNS SOA records, extract the hostmaster contact stored in the RNAME field as defined in RFC 1035 [17], and check whether the syntax of the email address is correct. Note that the domain name of an email gathered from the RNAME field may be different from the tested domain itself implying that we also need to lookup the MX record of the hostmaster domain.

Algorithm 1 formalizes the email validation procedure.

Algorithm 1 Email Address Validation with SMTP

```
1:  procedure SMTP(mx_server, addr)
2:      recipient email ← addr
3:      smtp_connection(mx_server)                ▷ on port 25
4:      if receive() = "220" then
5:          send("EHLO <sender hostname>")
6:          if receive() = "250" then
7:              send("MAILFROM: <sndr email>")
8:              if receive() = "250" then
9:                  send("RCPT TO: <rcpt email>")
10:                 if receive() = "250" then
11:                     return "valid"
12:                 else
13:                     return "not valid"
14:     return "error"
15: procedure reachable(mx_server, list_addresses)
16:     catchAll ← False
17:     for addr in list_addresses do             ▷ in parallel
18:         result[addr] ← SMTP(mx_server, addr)
19:     if result[random_addr] = "valid" then
20:         return "Catch All"
21:     else
22:         return "result[addr]" ▷ further parse result
```

"220": Simple mail transfer service ready
"250": Requested mail action okay, completed

2.1. Domain Sampling

To compare the results of the most popular websites with less known ones, we leverage top 1 M domains from Tranco—a domain ranking list oriented toward research, which uses the known ranking lists: Alexa, Cisco Umbrella, Majestic, and Quantcast, with additional improvements against ranking manipulation [12].

To measure the reachability rates of the different TLDs, we perform the experiment on generated samples of `.com` and `.net` gTLD names derived from zone files under the contract of VeriSign Inc, new gTLD names made available by the ICANN Centralized Zone Data Service[4], and ccTLD names. As most of the ccTLD registries do not make their zone files available to third parties, we leverage such domains from the Forward DNS data maintained by Rapid7[5]. We categorize the domains based on their TLD and we sample domains from each group.

We calculate the size of each sample using binomial approximation [18]. To estimate the positive rate p of each sample, we conduct a preliminary scan on randomly drawn domain samples of an arbitrary size (greater than 80,000 domains per sample). We define the positive rate p as the rate of domains in the CONTACT FOUND category, while the negative rate $1 - p$ is defined as the rate of active domains in the NO MX RECORD, NO CONTACT FOUND, and CONNECTION ERROR categories. The CATCH ALL category is considered neutral because we cannot determine if a domain is reachable or not. To circumvent such an exception, we split the neutral rate into positive and negative ones based on the proportion of the two previously computed rates. By doing so, we assume

[4]https://czds.icann.org
[5]https://opendata.rapid7.com/sonar.fdns_v2

TABLE 1. ESTIMATED FRACTION OF DOMAINS WITH VALID CONTACTS (POSITIVE RATE) AND SAMPLE SIZE

Parameters	Samples				
	.com	.net	ccTLD	ngTLD	AXFR
positive rate	0.043	0.053	0.067	0.015	0.068
sample size	633,663	772,338	966,867	232,845	975,276

that the CATCH ALL domains have the same positive rate as domains without CATCH ALL.

We finally set the confidence level to 95% and the standard error to $e = 0.001$. The sample size is then computed using the formula $n = 4 * \frac{1.96^2 * p(1-p)}{e^2}$. We randomly sample n domains from each population based on the estimated positive rates (see Table 1).

Furthermore, we identify the domains vulnerable to *1)* DNS AXFR transfers [13] (975,276 domains sampled from a list of 1,436,838 using the same procedure), *2)* DNS zone poisoning [14] (10,080 domains) (we followed the measurement methodology and ethical principles as described by the authors), *3)* web domains with known vulnerable WordPress versions (2,552 domains found using WPScan[6]), and *4)* compromised domains. We use a list of compromised domains collected from different public posts where malicious actors publish their hacking achievements. The list has 40,934 domains.

We perform our study on a total of 4,602,907 domains after removal of 31,648 duplicates present in more than one sample: e.g., `google.com` is present in the Tranco list and has been randomly drawn for the `.com` sample.

2.2. Ethical Considerations

We perform the scan of public SOA contacts and RFC abuse contacts—a total of 7 contacts per domain (including a randomly generated email address) but we do not send any emails. We deploy a website accessible from the IP address of the scanning server and the domain name of the sender (FROM contact) to allow them to opt out or ask for additional details of our measurements. We received one request to opt out from the study.

3. Results of Email Validation Scans

3.1. SOA vs. RFC Generic Email Reachability

The scan results are shown in Table 2. We consider a domain as reachable if at least one RFC-specific generic email alias has been validated. It is then labeled as CONTACT FOUND. We find that in all the studied categories, domains are more reachable using SOA contacts. For instance, 39.74% of Tranco top 1 M domains are reachable using an email leveraged from the SOA RNAME field while only 24.16% are reachable using RFC-specific contacts. We also find significantly more missing MX records for RFC-specific contacts than for mail servers of SOA contacts because DNS SOA records are often maintained by DNS service operators and less frequently by the domain owners who often do not have enough expertise in configuring DNS servers. Nevertheless, the SOA contact can be considered as a direct communication channel in case of, for example, misconfigured DNS servers vulnerable to AXFR transfers [13] or non-secure DNS dynamic updates allowing domain name hijacking [14].

[6]https://wpscan.org

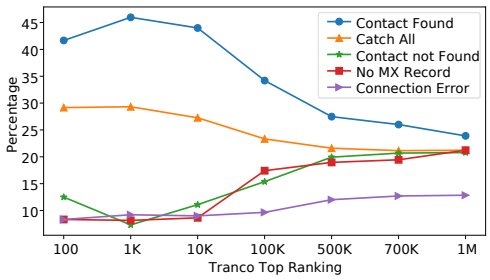

Figure 1. Results of validation scans of RFC-specific email aliases for the Tranco top 100 to Tranco top 1 M lists.

3.2. Tranco Top 1 M Popularity List

The results for the Tranco top 100 to 1 M lists (Figure 1) show that the better the rank of the domain is, the more likely the domain is reachable and complies with RFC 2142. For the top 1 K domains, at least 43.9% are reachable and the proportion gradually decreases to 22.6% for Tranco top 1 M. Similarly, the rate of CATCH ALL domains decreases from 28% (top 100) to 19% (top 1 M). The results can be explained by the fact that the operators of more popular websites put more emphasis on preventing email address enumeration of their clients.

We also observe an increasing rate of domains with invalid contacts: from 7% (top 1 K) to 19.7% (top 1 M) and domains without MX records: from 8% (top 1 K) to 20% (top 1 M). The connection error rate has also slightly increased from 7% (top 100) to 11.5% (top 1 M).

3.3. TLD Sampled Domains

Motivated by the expansion of the domain name space and the introduction of ICANN New gTLD Program, we now study the differences between country-code, legacy, and new generic TLDs.

The ccTLD names seem to be the most reachable for both SOA (35.4%) and RFC-specific (12.68%) contacts when compared to `.com` (20.20% and 8.97%) and especially to new gTLDs (21.35% and only 3.97%). New gTLDs names are far less reachable in the sample with 61.63% of domains without MX records, three times more than ccTLDs and almost twice more than `.com` domains.

The ccTLD market is more mature and registry operators are often non-profit organizations, so they do not have to compete aggressively and reduce registration prices [19]. They tend to invest more in security measures and research (e.g., REMED3IS [20], COMAR [21], PREMADOMA [22]). The new gTLDs are at the other end of the spectrum of the ecosystem. They are often funded by private investors, so their main focus is on revenues [23]. Security, and in particular, reducing domain names abuse, may not always be their primary objective [19]. Previous work showed increased abuse rates in some new gTLDs. For example, as many as 51.5%, 47.6%, and 33.4% of all registered domains in `.science`, `.stream`, `.study` TLDs, respectively, were blacklisted by Spamhaus in the 4th quarter of 2016 [19]. Apart from maliciously registered domains, many registrations are promotional, speculative, or defensive in nature [23]. Previous work showed that domains are often parked, do not resolve, or do not serve any meaningful content [19], [23]. Therefore, it is not

TABLE 2. RESULTS OF EMAIL VALIDATION SCAN ON THE SELECTED TLDS, VULNERABLE AND POPULAR DOMAINS

Category		Domains (%)								
		Selected TLDs				Selected vulnerabilities				Popular
		.com	.net	ccTLD	ngTLD	Comprom.	AXFR	Zone Poisoning	WP vuln.	Tranco 1 M
RFC emails	NO MX RECORD	35.48	35.59	18.43	61.63	25.95	20.23	39.80	21.01	21.46
	CONN. ERROR	25.58	27.32	19,11	9.68	15.37	18.51	9.84	12.86	13.00
	CATCH ALL	12.33	11.42	18.86	7.68	18.03	15.57	19.49	21.51	20.36
	NO CONTACT FOUND	17.63	17.80	30.91	17.03	28.26	31.08	18.40	27.34	21.02
	CONTACT FOUND	8.97	7.87	12.68	3.97	12.39	14.61	12.47	17.29	24.16
SOA contact	NO SOA RECORD	4.49	4.66	3.63	11.01	10.44	27.53	20.23	10.34	2.84
	NO MX FOR SOA	9.01	7.70	8.31	10.47	10.25	8.39	11.70	7.80	9.04
	CONN. ERROR	34.68	35.83	15.30	15.86	23.61	10.05	23.67	14.34	18,60
	CATCH ALL	18.09	17.54	22.36	12.12	17.45	15.49	16.61	14.89	19.38
	NO CONTACT FOUND	13.53	13.28	14.99	29.17	10.50	14.62	13.17	9.13	10.40
	CONTACT FOUND	20.20	20.97	35.40	21.35	27.75	23.91	14.62	43.38	39.74

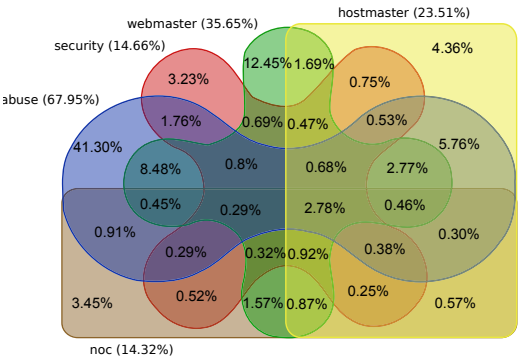

Figure 2. Venn diagram of the used RFC-specific email aliases

surprising that new gTLD domain names are also less reachable than ccTLD and legacy gTLD domains.

3.4. Vulnerable and Compromised Domains

We conclude that, in general, vulnerable and compromised domains are more reachable compared to the domains sampled from the general population. Interestingly, using email addresses stored in the RNAME field of DNS SOA records, domains with known vulnerable versions of WordPress are more reachable than Tranco top 1 M domains (43.38% versus 39.74%). We suspect that domains that use WordPress and other CMSs are hosted on a shared hosting infrastructure and therefore, it is usually the provider that maintains DNS SOA records.

3.5. Trends of Used RFC Names

We next briefly analyze the most common RFC-specific email aliases. Figure 2 presents the results in the Venn diagram. We observe that the most frequently used contact is *abuse* (67,95%). As many as 64.79% of domains only use one valid contact: 41.3% of domains use the *abuse* contact, while 23.46% use one of the other 4 aliases (half of them use *webmaster*). 35.21% of domains use multiple RFC-specific contacts.

4. Regression Analysis of Reachable Domains

To verify statistically different interpretations of the results, we build a Generalized Linear Model (GLM) using logistic regression. The dependent variable of the model is whether a domain is reachable. The independent variables are indicators that we measure and group in four categories: the owner and administrator efforts, the value of the domain, and whether the domain is on a shared host.

4.1. Data Collection

4.1.1. Efforts of domain owner. We assume that if the domain owner put effort to configure it properly, the probability of domain reachability also increases. We propose the following variable to indirectly measure the owner's effort:

DNSSEC deployment (has_DNSSEC): We check if the domain owner correctly deploys DNSSEC [24] to prevent DNS cache poisoning attacks. It can either be configured by the domain owner or deployed by a domain registrar or a reseller. Therefore, it generates indirect or direct costs to the owner. We first scan for the DNS RRSIG resource record (RR). If the record is present, we do a DNS A record lookup using a Google public resolver. If DNSSEC is correctly deployed, it responds with the IP of the domain.

4.1.2. Effort of the DNS administrator. The DNS administrator is responsible for setting a valid SOA RR in the zone file. To measure indirectly the effort of an administrator, we check whether the DNS zone of the domain is vulnerable to ***zone poisoning (poisoning_vuln)*** [14].

4.1.3. Value of the domain. We check if the value of the domain is a factor increasing its reachability. We propose the following variables:

Presence in Tranco top 5 M (tranco): If the domain is ranked in the Tranco popularity list [12], we expect it more valuable and therefore reachable.

Number of captures in the Internet Archive (log_archive_count): Internet Archive[7] is a non-profit organization collecting billions of captures of Internet websites. The higher the number of captures, the higher the value of the website.

Age of the first Internet Archive capture of the domain (log_first_seen): The feature is calculated by counting the days from the date of the first web capture of the domain in the Internet Archive till the measurement date.

Spamhaus badness index[8] (sp_index): It represents the

[7] www.archive.org

[8] https://www.spamhaus.org/statistics/tlds

abuse rate of individual TLDs. The more maliciously registered domains in relation to all active domains, the higher the TLD badness index and thus, the lower the value of the domain. For comparison, on May 12th, 2020, the `.fr` ccTLD has a badness index of 0.09 (1.4% of bad domains), while the `.top` new gTLD has an index of 3.84 (42.6% of bad domains).

4.1.4. Shared hosting/DNS services. Shared service providers maintain administrator privileges on their clients' websites and typically share responsibility for technical aspects of the servers with webmasters [25]. Domains on shared hosting are expected to be more reachable compared to domains using private or managed services. Shared hosting is measured via three variables:
Shared DNS server name (ns_shared): We first scan for DNS NS RR of each domain and extract the domain part of the fully qualified domain name (FQDN) of the authoritative name server using the public suffix list (for example, we 'normalize' `ns14.domaincontrol.com` to `domaincontrol.com`). We calculate the name server index as the proportion of domains that share the same 'normalized' name server to all domains in the sample. We assign to each domain the calculated NS index. The bigger the NS index, more reachable such a domain should be using the SOA contact considering that the DNS provider of such a popular name server correctly configures SOA RR.
Usage of a CMS (has_cms): Content management systems and e-commerce platforms are often proposed and maintained by hosting provider services. We detect the usage of popular CMSs such as WordPress, Drupal, or Joomla using Wappalyzer[9]. We expect that domains using CMS will rather be hosted on shared hosting, and therefore more reachable.
SPF record (has_spf): We check if there is an SPF rule in the DNS TXT record of each domain to prevent email spoofing [26]. Although, it can be configured by the owner of the domain, it requires significant expertise. Registered domains often have pre-configured SPF records (we confirmed this practice for large registrars such as GoDaddy, OVH, or Porkbun). Therefore, this variable indicates that the domain is managed by the service provider.

4.2. Results of Regression Analysis

We model the reachability of domains using RFC-specific and SOA contacts using logistic regression. The choice of the model comes from the binary nature of the dependent variable (reachable vs. not reachable). In Table 3, we observe two models with the same variables but with different dependent variables: RFC reachability in the left column and SOA reachability on the right. The final models in Table 3 are chosen based on a stepwise addition of the variables into a baseline model with a single explanatory variable. The *log_first_seen* and *log_archive_count* values are transformed with the *log* function to make them follow a normal distribution. After removing domains with missing values, we build the model based on 2,837,322 observations for RFC reachability and 2,711,165 observations for SOA.

As expected, the coefficient *has_cms* is positive in both models (RFC and SOA) meaning that domains using CMS are more likely to be reachable than domains without it. While holding all other variables constant, the probability of reaching a domain with RFC-specific contact increases from 7.99% (for a domain that does not use a CMS) to 8.76% (a domain with CMS).

TABLE 3. RESULTS OF TWO LOGISTIC REGRESSION MODELS

	Dependent variable:	
	reachable_rfc	reachable_soa
	(1)	(2)
has_cms	0.100*** (0.004)	0.150*** (0.003)
ns_shared	0.023*** (0.0004)	0.365*** (0.001)
has_spf	0.457*** (0.003)	0.139*** (0.003)
has_dnssec	0.035*** (0.011)	0.015* (0.009)
sp_index	−0.397*** (0.006)	−0.459*** (0.004)
tranco	0.278*** (0.005)	0.164*** (0.004)
log_first_seen	−0.008*** (0.001)	0.002*** (0.001)
log_archive_count	0.219*** (0.001)	0.053*** (0.001)
poisoning_vul	−0.384*** (0.058)	−1.004*** (0.051)
Constant	−2.443*** (0.005)	−0.795*** (0.004)
Observations	2,837,322	2,711,165
Log Likelihood	−1,239,917.000	−1,711,767.000
Akaike Inf. Crit.	2,479,854.000	3,423,554.000
Note:	*p<0.1; **p<0.05; ***p<0.01	

Moreover, the *ns_shared* variable shows that the higher the number of domains in the same name server, more reachable the domains are, which justifies the hypothesis that larger DNS providers are likely to have better DNS configurations and reachable SOA contacts. As expected, the SOA *ns_shared* coefficient is more significant than the coefficient for RFC by a factor of 16.

Interestingly, the fact that the owner has configured DNSSEC seems to be less relevant to whether or not the domain is reachable.

By looking at the coefficients of the Tranco top 5 M, the Spamhaus badness index (negative values of coefficients), and the number of captures of Internet Archive, we conclude that more valuable domains are also more reachable. The age of the domain does not appear relevant to reachability.

Domains vulnerable to non-secure dynamic updates (i.e., zone poisoning) are less likely to be reachable. The high negative coefficient observed for the SOA contact is to be expected as invalid SOA records and non-secure dynamic updates are both DNS misconfigurations.

Finally, we evaluate the highest probability for a domain to be reachable. We consider a high-value domain without vulnerabilities, for which the DNS service is operated by a large DNS provider such as Domaincontrol that covers 13% of the sampled domains. The website uses CMS and has an SPF rule, which indicates shared hosting. We assume that we observed 1100 Internet Archive captures. We find that such a domain would have a probability of 53.7% to have at least one RFC-specific email alias configured correctly and 99% to be reachable with the email stored in the RNAME field of SOA RR. This difference between the two probabilities is determined by the difference in the coefficients of *ns_shared* in both SOA and RFC models. The results show that SOA contact proves to be the most valuable when it is managed by large DNS providers.

[9]https://www.wappalyzer.com

4.3. Proposed Notification Scheme

The Internet community needs a large-scale notification system that covers the largest part of Internet domains for direct reporting of threats or abuses. Registrants are already required to provide a valid email address to registrars. Therefore, ICANN could oblige all accredited registrars to maintain an RFC-specific email alias (e.g., *abuse* as it is the most commonly used) for domains with a default MX record and redirect security notifications to mailboxes of domain owners. However, such a solution would not only lower the barriers in sending automatic notifications but also in sending spam and phishing emails at scale. To mitigate this effect, the community could define a standardized format of notifications to send on such channels to allow for automatic discarding of spam emails.

5. Conclusion

With the increasing effectiveness of large-scale vulnerability scanning, our results show the absence of an equally effective large-scale notification system. Previous studies showed that direct communication channels are very effective towards fast fixes. However, obtaining contact information at scale is highly problematic and therefore not suitable for large-scale notifications, especially after GDPR. We find that the top-ranked domains are the ones that follow the most RFC specifications. So, they are less sensible to the absence of a large-scale notification system because they are better secured.

We have used Logistic Regression to model and exhibit the relationship between domain reachability and different variables that reflect the owner and administrator effort, the value of the domain, and if the domain is on a shared host or not. We have shown that the probability of a domain having a valid contact increases if it is valuable or hosted on large shared hosting platforms (implying better DNS configurations).

Finally, we have proposed a possible scheme for large-scale notifications that does not require engaging millions of registrants. Instead, registrars could maintain email aliases for domains with a default MX record and redirect security notifications to personal email addresses of domain owners.

Acknowledgments

This work has been carried out in the framework of the PrevDDoS project funded by the IDEX Université Grenoble Alpes IRS and partially supported by the Grenoble Alpes Cybersecurity Institute CYBER@ALPS under contract ANR-15-IDEX-02, PERSYVAL-Lab under contract ANR-11-LABX-0025-01, and DiNS under contract ANR-19-CE25-0009-01.

References

[1] F. Li, Z. Durumeric, J. Czyz, M. Karami, M. Bailey, D. McCoy, S. Savage, and V. Paxson, "You've Got Vulnerability: Exploring Effective Vulnerability Notifications," in *USENIX Security*, 2016.

[2] O. Cetin, C. Gañán, M. Korczyński, and M. van Eeten, "Make Notifications Great Again: Learning How to Notify in the Age of Large-Scale Vulnerability Scanning," in *WEIS*, 2017.

[3] Z. Durumeric, J. Kasten, D. Adrian, J. A. Halderman, M. Bailey, F. Li, N. Weaver, J. Amann, J. Beekman, M. Payer, and V. Paxson, "The Matter of Heartbleed," in *Proc. IMC*, 2014.

[4] F. Li, G. Ho, E. Kuan, Y. Niu, L. Ballard, K. Thomas, E. Bursztein, and V. Paxson, "Remedying Web Hijacking: Notification Effectiveness and Webmaster Comprehension," in *WWW*, 2016.

[5] B. Stock, G. Pellegrino, C. Rossow, M. Johns, and M. Backes, "Hey, You Have a Problem: On the Feasibility of Large-scale Web Vulnerability Notification," in *USENIX Security*, 2016.

[6] B. Stock, G. Pellegrino, F. Li, M. Backes, and C. Rossow, "Didn't You Hear Me?—Towards More Successful Web Vulnerability Notifications," in *NDSS*, 2018.

[7] E. Zeng, F. Li, E. Stark, A. P. Felt, and P. Tabriz, "Fixing HTTPS Misconfigurations at Scale: An Experiment with Security Notifications," in *WEIS*, 2019.

[8] O. Çetin, M. H. Jhaveri, C. Gañán, M. van Eeten, and T. Moore, "Understanding the Role of Sender Reputation in Abuse Reporting and Cleanup," *J. Cybersecur.*, vol. 2, no. 1, pp. 83–98, 2016.

[9] "Temporary Specification for gTLD Registration Data," https://www.icann.org/resources/pages/gtld-registration-data-specs-en.

[10] "Advisory Statement: Temporary Specification for gTLD Registration Data," https://www.icann.org/en/system/files/files/advisory-statement-gtld-registration-data-specs-17may18-en.pdf.

[11] D. Crocker, "Mailbox names for common services, roles and functions," RFC 2142, 1997.

[12] V. L. Pochat, T. Van Goethem, S. Tajalizadehkhoob, M. Korczyński, and W. Joosen, "Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation," in *NDSS*, 2019.

[13] M. Skwarek, M. Korczyński, W. Mazurczyk, and A. Duda, "Characterizing Vulnerability of DNS AXFR Transfers with Global-Scale Scanning," in *IEEE Security and Privacy Workshops*, 2019.

[14] M. Korczyński, M. Król, and M. van Eeten, "Zone Poisoning: The How and Where of Non-Secure DNS Dynamic Updates," in *IMC*, 2016.

[15] J. Klensin, "Simple Mail Transfer Protocol," RFC 5321, 2008.

[16] Postfix Virtual Domain Hosting Howto. [Online]. Available: http://www.postfix.org/VIRTUAL_README.html

[17] P. Mockapetris, "Domain Names - Implementation and Specification," RFC 1035, 1987.

[18] E. Rahme and L. Joseph, "Exact Sample Size Determination for Binomial Experiments," *Journal of statistical planning and inference*, vol. 66, no. 1, pp. 83–93, 1998.

[19] M. Korczyński, M. Wullink, S. Tajalizadehkhoob, G. C. Moura, A. Noroozian, D. Bagley, and C. Hesselman, "Cybercrime After the Sunrise: A Statistical Analysis of DNS Abuse in New gTLDs," in *ACM AsiaCCS*, 2018.

[20] M. Korczyński, S. Tajalizadehkhoob, A. Noroozian, M. Wullink, C. Hesselman, and M. van Eeten, "Reputation Metrics Design to Improve Intermediary Incentives for Security of TLDs," in *Euro S&P*, 2017.

[21] S. Maroofi, M. Korczyński, C. Hesselman, B. Ampeau, and A. Duda, "COMAR: Classification of Compromised versus Maliciously Registered Domains," in *Euro S&P*, 2020.

[22] J. Spooren, T. Vissers, P. Janssen, W. Joosen, and L. Desmet, "PREMADOMA: An Operational Solution for DNS Registries to Prevent Malicious Domain Registrations," in *ACSAC*, 2019.

[23] T. Halvorson, M. F. Der, I. Foster, S. Savage, L. K. Saul, and G. M. Voelker, "From .Academy to .Zone: An Analysis of the New TLD Land Rush," in *IMC*, 2015.

[24] O. Kolkman and R. Gieben, "DNSSEC Operational Practices," RFC 4641, 2006.

[25] S. Tajalizadehkhoob, T. van Goethem, M. Korczyński, A. Noroozian, R. Böhme, T. Moore, W. Joosen, and M. van Eeten, "Herding Vulnerable Cats: A Statistical Approach to Disentangle Joint Responsibility for Web Security in Shared Hosting," in *ACM CCS*, 2017.

[26] S. Kitterman, "Sender Policy Framework (SPF) for Authorizing Use of Domains in Email," RFC 7208, 2014.