Feasibility of Large-scale vulnerability notifications after the GDPR

Wissem Soussi
Grenoble-Alpes University
Saint-Martin-d'Heres, France
me@wsoussi.com

ABSTRACT

In the last years many large-scale vulnerability detection tools are publicly available and open source. They are being used by researchers for their studies but even more by evildoers with malicious purposes. These acts are damaging the internet stability which is today a pillar in our economy, politic and social life. Therefore our concern is to find a large-scale notification system in order to reduce the exposure of domains with such security flaws. This system, possibly built on top of the existing communication channels and infrastructures, should be as effective as the tools of vulnerability detection. The mostly used direct communication channel for this purpose is the WHOIS domain contact. But a problem arises with the 2018 European law on data protection and privacy (the GDPR): domain owner's personal information is no more available in WHOIS, including its contact. In this study, we highlight the influence of the GDPR on all the studies done on vulnerability notifications. We propose a first methodology to obtain a valid email contact for a specific domain by systematically testing different communication channels with administrators and owners of the domains. Finally, we test this methodology over different domain samples. We give the results obtained in the study of different Top Level Domains (TLDs), compromised domains and vulnerable domains, showing the problem on the feasibility of large-scale vulnerability notification. Indeed, the rate of reachable domains is very low. Finally, we propose a possible solution to the internet community, in particular to the ICANN and to the registrars, in order to solve this problem.

KEYWORDS

GDPR, vulnerability notifications, WHOIS, SMTP, RFC, TLDs

1 INTRODUCTION

Large-scale vulnerability detection becomes widely used with scan tools for auditing network/internet security, like NMAP[12], ZMAP[8] and WPScan[4]. As some domain abuses should be directly notified to the interested domain administrators, the Internet community needs to maintain the ability of large-scale notification mechanisms. Thus, to directly inform the affected parties of the vulnerability. Internet community has set up indirect communication channels using reliable third entities such as CERTs and registrars which validate the notification as a detection of a threat and notify in turn domain administrators. This process may take time and in our case study, we want to test the reachability of direct communication channels, where usually the vulnerability fix is done shortly after its identification.

In May 25 2018, the General Data Protection Regulation (GDPR)[1] came into effect. This requires companies to get consent for personal information they gather on EU citizens and residents. IN order to be conform to the new law, the Internet Corporation for Assigned Names and Numbers (ICANN) adopted the Temporary Specification for gTLD Registration Data [3] and some registration information is no more displayed in the public WHOIS data, in particular, the Registrant and Administrative Contact. Therefore, the new regulation is further reducing the effectiveness and feasibility of domain vulnerability and abuse notifications. In case the registrant contact is absent, normal practice is to contact the related registrar, who is required to respond in "a reasonable timeâĂİ. But for the same reasons of CERTs, this may introduce important delays on the fixing rates.

Several researchers have begun investigating how security notifications should be sent in order to increase the rate of revision and vulnerability fixing [6, 10, 11, 15–17]. Frank Li et al. [10] for example studied the impact of the elements composing a notification: content verbosity, external redirection links, message language translation and they have confronted effectiveness of notifications with CERTs as intermediaries against direct notifications sent using WHOIS contacts. And indeed, results show that domains notified with WHOIS contacts had a faster and greater fixing rates.

This logically implies that GDPR has an impact on today's large-scale vulnerability notifications, and renews this topic as an actuality. Moreover, no one has systematically studied a more fundamental problem, which is the reachability rate of notifications over the internet as a global infrastructure, regardless to wether domains have vulnerabilities or not. Studies on notifications until now have always experimented reachability of domains over specific vulnerabilities. They have indeed get deeper results about notification effectiveness and fixing rates that we could not have otherwise, but a less great picture of the internet situation overall. This is the main concern and contribution of our work.

We systematically test different direct communication channels in order to quantify the rate of domains that are effectively reachables. We test direct contacts like generic email addresses enumerated by the RFC 2142 [7]: for a domain *example.com*, RFC requires valid email adresses such as *abuse@example.com* or *security@example.com* for abuse notifications. We estimate the rates of RFC generic email addresses that are correctly configured, reachable, and for which notifications may be successfully delivered. We also test Start of Authority (SOA) contacts, gathered from DNS lookups and related to the hostmaster of the domain.

We perform the study on a higher number of domains than the other studies about large-scale notifications: 4,317,428 domains including Alexa top 1 million domains and representative samples

Table 1: Previous studies on notifications before GDPR

Reference	tested channels	vulnerabilities	# domains
Stock et al. 1 [16]	WHOIS, generic RFC, CERTs and hosting providers	WP, r-XSS and cs-XSS	35,832
Stock et al. 2 [15]	mail, social networks, phone, WHOIS and generic RFC	public git reps and WP	24,000
Li et al. [11]	Chrome browser alerts, Google Search Console and WHOIS	Hijacked websites	760,935
Li et al. [10]	WHOIS and CERTs	ICS, IPv6 firewall and amplifiers	310,227
Cetin et al. [6]	WHOIS, SOA and generic RFC	Zone Poisoning	21,506
Zeng et al. [17]	Google search console and WHOIS	HTTPS misconfigs	44,548

of different TLD's domains such as gTLD *.com* and *.net* domains, ccTLDs domains, new gTLDs domains, and also more targeted domains such as domains using WordPress, compromised domains and vulnerable domains. The vulnerabilities are three, two of them are recent and related to DNS security: AXFR transfers[14] and zone Poisoning[9]. The third is more global and uses the WPScan tool to find WordPress[4] vulnerabilities.

Intuitively, in our research study the notification *per se* is not a concern, since we do not only look for vulnerable domains. Indeed, we do not send vulnerability notifications to the targetted domains. We do not want to send useless messages and spams to domains for the seak of our study, and this for a simple ethical reason. We design a scanning tool that checks the existence and validity of an email address by avoiding to actually send emails.

In the representative sample of *.com* domains, by applying the designed methodology (Section 4), we show that at least 71% of the domains are not reachable using any RFC standard name, in particular because they do not have MX records (56%). within domains with valid MX records we are able to confirm that 13% have at least one valid contact. Interestingly, the most used RFC generic address is in absolute '*abuse*' (94%), followed by '*webmaster*' (15%) and 22% of all the reachable domains have more than one valid generic contact. For SOA records at least 23% of *.com* domains have valid contact.

2 RELATED WORK

As we can notice in Table 1, all studies performed on vulnerability notifications used WHOIS contacts. For direct communication with domain owners Google has also used a private communication channel such as Google Search Console, which was effective but covered only part of the targeted vulnerable domains. Stock et al. [15] studied alternatives methods like physical mails, social media contacts or phone calls, which do not suit large-scale notifications. However, in another study Stock et al. [16] tested more suitable communication channels such as generic RFC contacts, WHOIS contacts and indirect channels with CERTs and hosting providers as intermediaries. They found interesting results (similar in Cetin et al. [6]) showing how large-scale notifications can be effective on the fix rate of domains who received the notification, but how only 5.8% of the targetted domains actually received the notifications. These results can be confirmed with our current study, showing low reachability rate for the most used gTLDs domains such as *.com*, *.net* and even worst results in new gTLDs. Alexa top 1 million ranked domains had a reachability of 20.6%, which is to be considered low for the set of the most sought after domains.

3 BACKGROUND

The domain contact gathered from the DNS SOA (Start of Authority) is available in the field RNAME as defined in the RFC 1035 [13]. This contact is the email address of the responsible person for the domain name zone. In few cases, when the DNS zone of the domain is managed by the domain owner, this contact is a direct communication channel. However, in most cases domains rely on DNS service providers which put their contact in the SOA RNAME as the DNS zone administrators. We then classify the SOA contact as indirect, but we still include it to our scanning system.

Another information we can gather from DNS is the exchange mail servers of a given domain, saved in the DNS zone file as MX records [2]. These servers are publicly interfaced to the Internet in order to receive emails from other mail servers sent toward internal mailboxes.

4 SYSTEM DESIGN

These mail servers communicate using the SMTP (Simple Mail Transfer Protocol), standard defined by the RFC 5321[2]. Communication happens through a TCP connection. SMTP can also be used on top of an encrypted connection (mainly with STARTTLS) or can require client authentication. However client authentication is required when the mail server is used as a relay point, to redirect emails toward other external mail servers. This restriction avoids the abuse of the mail server relays for sending spamming or phishing emails by also spoofing the IP address of the sender. MX servers instead are used to receive emails forwarded to internal addresses and here authentication is not required. Regarding encrypted connections, this is very rarely imposed by MX servers but receivers are notified about the email being in plain text.

During the procedure of sending an email, with the 'RCPT TO' query, we are able to verify if the receiver's address exists without actually sending any email. However, for each mail server this procedure should be used only few times, since some mechanisms against email address-mapping may close the communication and blacklist the client's IP. Address-mapping allows to get a list of valid email addresses to use for malicious purposes. For this reason different countermeasures are adopted by some mail servers. One of them is to allow only a limited number of 'RCPT TO' queries from the same TCP communication. Another countermeasure is to accept any given address, even wrong ones. We note this procedure 'Catch All'. This is inconvenient for us, since we are not able to confirm the validity of an email address when the mail server adopts this procedure.

We develop a scanning software tool that takes as input a list of domains and replies for each domain whether it is reachable,

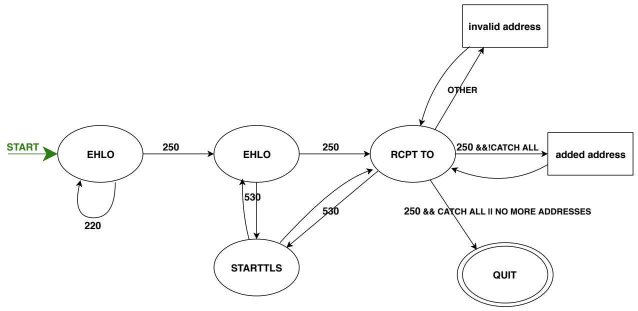

Figure 1: Algorithm for email address validation

meaning it has at least one valid contact, or not. DNS entries like SOA and MX records are dynamically gathered at every scan, and within the email addresses to check we also check a random address to verify whether the mail server adopts the 'Catch All' or not.

The first step is to get the list of contacts we aim to validate for each domain:

- we take the hostmaster contact in the RNAME field of the Start of Authority (SOA) after checking the syntax is correct.
- we generate email addresses using the names enumerated in RFC 2142 [7]: e.g. for the domain *example.com* we would have *hostmaster@example.com* and *webmaster@example.com* for DNS and HTTP issues, *abuse@example.com*, *noc@example.com*, *trouble@example.com* and *security@example.com* for network abuse and vulnerability notifications.

The next step is to connect, for each domain, to the mail server with the highest priority present in the MX record, and to communicate with it using SMTP with a defined "conversation" represented in figure 1 and showing a sequence of replies to send based on answers we get from the mail server. However, instead of verifying first the catch all random email address and then the other addresses, they are all verified in parallel each one on a different communication, this in order to avoid the limitation of the mail server of using 'RCPT TO' query a limited number of times in a single connection. The SOA contact domain may be different from the targeted domain itself, implying we would need to lookup another domain's DNS zonefile to get the MX record of the SOA contact.

4.1 Output

The implemented program used for the tests does not switch to a secure connection when the MX server requires it. This because MX servers almost never require secure connection (confirmed during tests).

The results are elaborated by the program with respect to domains, which are classified according to 10 exclusive and complete categories (a domain belongs to at least one and only one category):

(1) Inactive: the domain has no A or AAAA records.
(2) No MX record: the domain doesn't have mail servers.
(3) Catch All: the mail server is accepting all addresses.
(4) Contact Found: at least one contact has been validated.
(5) No Contact Found: no contact has been validated.
(6) Server Unreachable: the MX record is present but it is either invalid, either the mail server is not found.

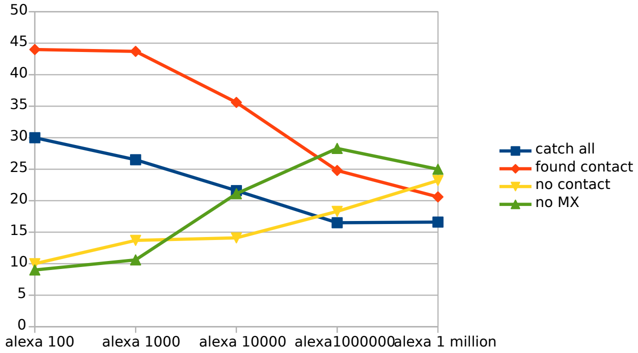

Figure 2: Alexa progressive rate course

The other categories are related to scan failures like Timeout (7), SMTP Communication Error (8), SMTP Communication Refused (9) and TLS Required (10).

5 TESTS

5.1 Domain sampling

We want to know how different the results may be by taking the top ranked websites in comparison to less ranked ones and finally give approximation rates of reachability on the *.com* general top level domain (gTLD) using a sample of domains. For domain ranking based on web traffic we use Alexa top 1 million.

We also apply the test on generated samples of *.net gTLD*, *ccTLDs*, *newTLDs* and on a public list of compromised domains taken from *zoneh.org* [5]. Each generated sample is randomly drawn from zonefiles of hundreds of millions domains representing all domains the targeted gTLDs. zonefiles for .com GTLD, *.net GTLD* and *new GTLDs* are shared with us under the contracts with *Verisign* and *ICANN*. The *ccTLDs* are gathered from the public database of The goal in the sampling process is to have a good representativeness of the results obtained on the sample with respect to what we would obtain by inspecting the total population under analysis. For each population under study we list all domains, each in a single line. Afterward we randomly choose a line between the first and the last one line and we drawn the domain name for the sample. We iterate the process until we have the adequate sample size.

We determine the sample size statistically by taking into consideration 3 factors:

- **The Variance** is the width of the range of variation on the results of the experiment. The bigger is the variance, the bigger is the sample size to take. Its value is not easy to determine before the tests. This is why we conduct a pilot study to define the expected variance. We use a scanner which follows the same principles of the designed system, but without using a mail server. This implies a bad reputation level of the sender since the IP used is do not have a related public domain name and this translates to a not conformity with the Sender Policy Framework (SPF). also DNSBL organizations will not be informed of this first test, increasing probabilities to get blacklisted by the high number of SPF-non-conform

Table 2: Results of the scans without mail server

category/domains	**A 1m**	**.com**	**.net**	**ccTLD**	**new gTLD**	**compromised**	**AXFR**	**Zone Poisoning**	**WP**	**WP vul**
NOT REACHABLE	60.2%	90.4%	85.3%	81.9%	92.5%	82.5%	83.3%	84.3%	69.5%	80.7%
$\rightarrow NoMXrecord$	18.9%	55.7%	27.2%	35.7%	32.1%	23.4%	13.2%	26.4%	24.9%	13.7%
$\rightarrow Nocontactfound$	23.2%	14.8%	15.5%	24.7%	9%	17.1%	24.9%	19.3%	33.7%	22.6%
$\rightarrow Serverunreachable$	0.6%	0.9%	0.7%	0.9%	0.4%	0.4%	1.8%	0.9%	0.6%	0.2%
$\rightarrow Others$	14.4%	8.8%	29.2%	20.2%	12.9%	39.8%	24.7%	14.5%	10.3%	43.5%
REACHABLE	20.6%	4.1%	5%	6%	1.5%	6.1%	6.5%	4.9%	14.9%	9.6%
CATCH ALL	16.6%	4.3%	6.4%	11.9%	2.4%	10.7%	4.9%	5.6%	15.5%	9.3%
VALID SOA	28.8%	22.9%	22.8%	22.2%	19%	18.3%	15.4%	8.5%	47.5%	52%

Table 3: Results on the different domain samples

category/domains	**.com**	**.net**	**ccTLD**	**new gTLD**
positive rate	0.056	0.064	0.061	0.029
variance	2.044E-5	2.341E-5	1.461E-5	1.126E-5
standard error	0.0045	0.0048	0.0038	0.0034
confidence width	0.0089	0.0095	0.0075	0.0066
sample size	260,488	257,412	390,078	248,255

‘mail from‘ requests. However this first test is useful to have the next results statistically as precise as possible. Results of the first tests are present in table 2 and discussed with the final results in section 5.2 . After retrieving the positive rate corresponding reachable domains, noted p, the variance is calculated as $variance(p) = \frac{p(1-p)}{n}$ where n is the number of the scanned domains.

- **The standard error** represents the variability index of our statistics and is $standarderror(p) = \sqrt{variance(p)}$. The variability of the positive rate is then equal to $p \pm standarderror(p)$.
- **The Confidence Level** is used on the standard error in order to calculate the confidence width. This basically increases the interval gave by the standard error by multiplying it to a certain value. By convention scientific studies use 90%, 95% or 99% of confidence level. Motivations for the choice are usually more practical than statistical, such as available resources (financial budget, population size, time etc..). Applying a confidence level of 95% we obtain a $confidencewidth(p) = 1.96 * standarderror(p)$.

Now using the same confidence level of 95% the sample size is calculated as $n = \frac{1.96^2 p(1-p)}{confidencewidth(p)^2}$. In table 3 we find the size calculated for each sample.

5.2 Results

The percentage of valid SOA contacts in table 2 is complementary and not exclusive with respect to the other ten categories defined above. A first observation we do is the similar validity rate of SOA contacts for all the tested samples, if compared to rates of other categories. Alexa top 1 million has the highest reachable SOA rate of 29%, followed by *.com*, *.net* and *ccTLDs* of around 23%. For *new gTLDs* and *vulnerable* domains rates lightly drop to 18-19%. ‘Catch All’ rates for SOA contacts are also similar (around 14%). With respect to reachability of domains with generic RFC contacts SOA reachability rate is consistently higher. This is justified by the fact that SOA is in most cases a third party contact added not by the domain owner but by the registrar or hosting provider.

The results on the Alexa 1 million domains (figure 2) show clearly that the better a domain is ranked the more probabilities it has to be reachable and to follow RFC specifications. Domain reachability is related to the rate of domains in the ‘Contact Found’ category. For the top 100 domains at least 44% of domains are reachable, and this gradually decreases to 20.6% by extending the sample to the less ranked domains. In the same way the ‘Catch All’ category decreases from 30% (top 100) to 16.6% (top 1 million). This can be interpreted as a reduction of scrupulosity of less ranked domain’s mail servers.

We notice instead an increasing rate of domains without valid contacts, from 10% (top 100) to 23.2%(top 1 million) and a similar variation on the rate of domains without MX records: from 9% (top 100) to 25% (top 1 million).

Moving to the other samples, where domains are not ranked (or at least not necessarily), we observe that domains reachability rate is very low (for all the samples the maximum rate is 6.1%). The domains of *new gTLDs* have a reachability rate of only 1.5%, which is justified compared to the results of other more popular TLDs like *ccTLDs*, *.com* and *.net* domains. *New gTLDs* domains are also the most ones which do not use the ‘Catch All’ countermeasure against address-mapping. In every sample, a considerable number of domains do not have MX records, especially for *.com* domains (55.7%) and *ccTLDs* domains (35.7%).

We take reachable domains of all samples and analyze the RFC names that are validated (used) the most (figure 3). RFC name *trouble* is neglected since only 2 domains over 80970 use it. As we can observe *abuse* is widely the most used name (83%). 66.65% of domains has only one valid RFC contact: 52.65% of domains has only the *abuse* contact while 14% has one of the other 4 contacts (3/4th of them use *webmaster* contact). The other domains (33.35%) has multiple valid contacts.

6 CONCLUSION

With the increasing effectiveness of large-scale vulnerability scanning, our result shows the inexistence of an equally effective large-scale notification system. Popular and top ranked domains are the ones who follow the most RFC specifications. But these domains can also be considered the ones who less need the large-scale notification system do to their better than average security. This brings to the conclusion that human factor is relatively important and not all domain owners can think to make available a vulnerability notification channel. For a solution to this problem we should

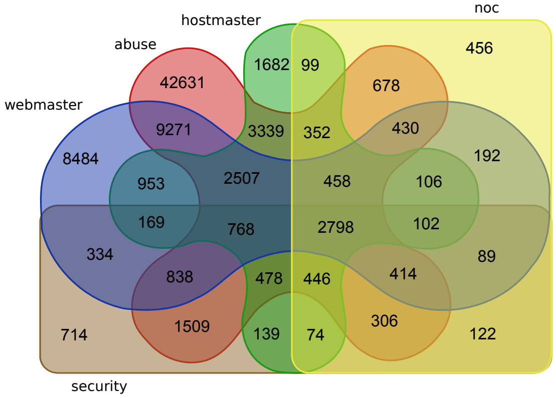

Figure 3: Venn Diagram of the used RFC names

not rely on registrants, which are hundreds of millions. Indeed we think that the problem can be solved by the main internet actors regardless to domain owner actions.

6.1 Possible Solution

We want to be conform to the GDPR, which means the private email address that domain owners give to registrars should not be public (unless the domain owner gives consent). Our main concern is to have a large-scale vulnerability notification system for the internet community with possibly a total coverage of all the public domains.

The solution we propose is for the ICANN to give the following requirements to all the registrars: the acquisition of a domain name urges the registrant to input a valid email address, verified by the registrar with an email allowing the domain activation. This is probably already performed by all registrars. Next, registrars will use a new or an already existing popular RFC generic address (e.g. *abuse@*) to redirect the emails toward the private mailbox without publishing its address. A similar concept was used by registrars before GDPR: when domain owners required data privacy, a randomly generated email address was published in the WHOIS contact. Emails sent to the generated address where then redirected by the registrar to the real address. However, this was not a mandatory task and for a large-scale notification system we needed a scraper to gather the random address from the WHOIS. Instead, using the RFC specification allows notifiers to directly have a known valid contact.

REFERENCES

[1] [n. d.]. The EU General Data Protection Regulation (GDPR). https://eugdpr.org/the-regulation/. ([n. d.]). Accessed: 2019.06.7.
[2] [n. d.]. SMTP. https://tools.ietf.org/html/rfc5321. ([n. d.]). Accessed: 2019-04-21.
[3] [n. d.]. Temporary Specification for gTLD Resgistration Data. https://www.icann.org/resources/pages/gtld-registration-data-specs-en. ([n. d.]). Accessed: 2019.01.27.
[4] [n. d.]. WPSCAN. https://wpscan.org/. ([n. d.]). Accessed: 2019-04-21.
[5] [n. d.]. ZONEH. https://zone-h.org. ([n. d.]). Accessed: 2019-07-04.
[6] Orcun Cetin, Carlos Ganan, Maciej KorczyÅĎski, and Michel van Eeten. 2016. Make Notifications Great Again: Learning How toNotify in the Age of Large-Scale VulnerabilityScanning. (2016).
[7] D. Crocker. [n. d.]. RFC 2142. https://tools.ietf.org/html/rfc2142. ([n. d.]). Accessed: 2019-04-21.
[8] Zakir Durumeric, Eric Wustrow, and J. Alex Halderman. 2013. ZMap: Fast Internet-wide Scanning and Its Security Applications. (2013).
[9] Maciej Korczyński, MichałKról, and Michel van Eeten. 2016. Zone Poisoning: The How and Where of Non-Secure DNS Dynamic Updates. In *Proceedings of the 2016 Internet Measurement Conference (IMC '16)*. ACM, New York, NY, USA, 271–278. https://doi.org/10.1145/2987443.2987477
[10] Frank Li, Zakir Durumeric, Jakub Czyz, Mohammad Karami, Michael Bailey, Damon McCoy, Stefan Savage, and Vern Paxso. 2016. You've got vulnerability: Exploring effective vulnerability notifications. (2016).
[11] Frank Li, Grant Ho, Eric Kuan, Yuan Niu, Lucas Ballard, Kurt Thomas, Elie Bursztein, and Vern Paxso. 2016. Remedying Web Hijacking: NotificationEffectiveness and Webmaster Comprehension. http://www.icir.org/vern/papers/notification-www16.pdf. (2016).
[12] Gordon Fyodor Lyon. 2009. *Nmap Network Scanning: The Official Nmap Project Guide to Network Discovery and Security Scanning.*
[13] P. Mockapetris. [n. d.]. RFC 1035. https://www.ietf.org/rfc/rfc1035.txt. ([n. d.]). Accessed: 2019-04-21.
[14] Marcin Skwarek, Maciej KorczyÅĎski, and Andrzej Duda. 2019. Characterizing Vulnerability of DNS AXFRTransfers with Global-Scale Scanning. (2019).
[15] Ben Stock, Giancarlo Pellegrino, Frank Li, Michael Backes, and Christian Rossow. 2018. DidnâĂŹt You Hear Me? âĂŤ Towards More SuccessfulWeb Vulnerability Notifications. https://publications.cispa.saarland/1190/1/stock2018notification.pdf. (2018).
[16] Ben Stock, Giancarlo Pellegrino, Christian Rossow, Martin Johns, and Michael Backes. 2016. Hey, You Have a Problem: On the Feasibility of Large-Scale Web Vulnerability Notification. (2016).
[17] Eric Zeng, Frank Li, Emily Stark, Adrienne Porter Felt, and Parisa Tabriz. 2016. Fixing HTTPS Misconfigurations at Scale:An Experiment with Security Notification. https://storage.googleapis.com/pub-tools-public-publication-data/pdf/06a75f932595f27a60092007965934c957b5de21.pdf. (2016).

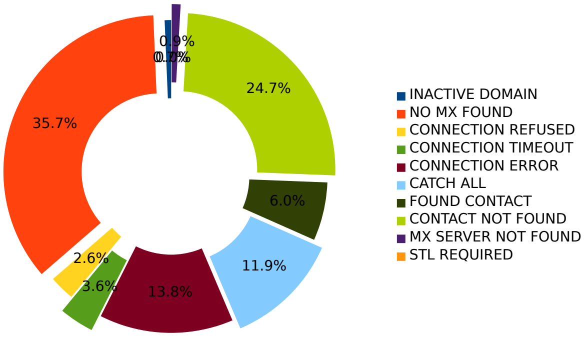

Figure 7: *ccTLD* sample results

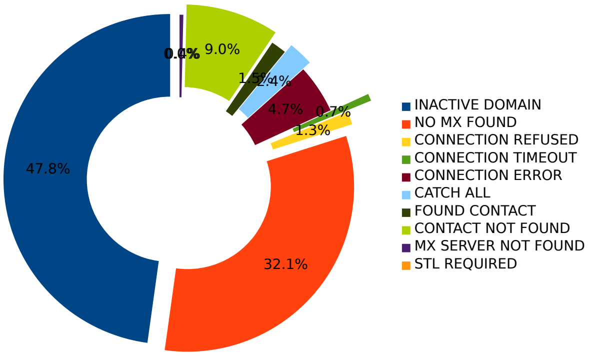

Figure 8: *new TLD* sample results

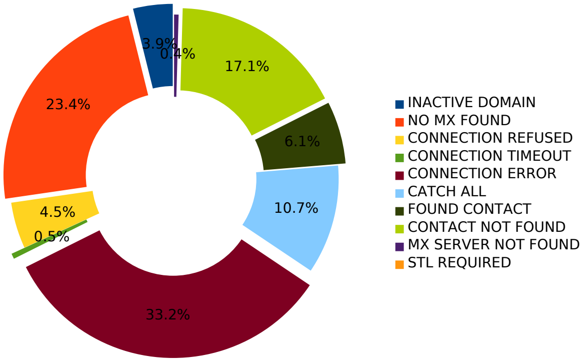

Figure 9: *new TLD* compromised domains from zone-h.org results

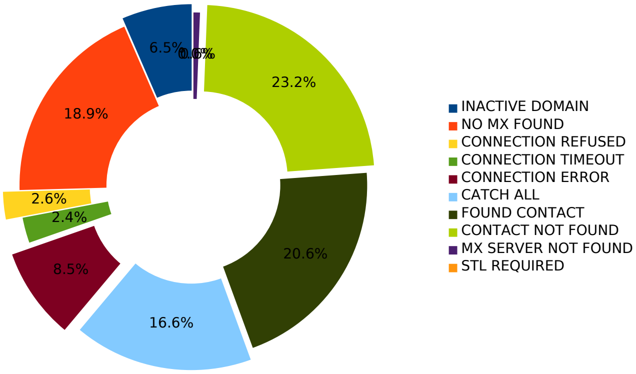

Figure 4: *ccTLD* alexa top 1 million results

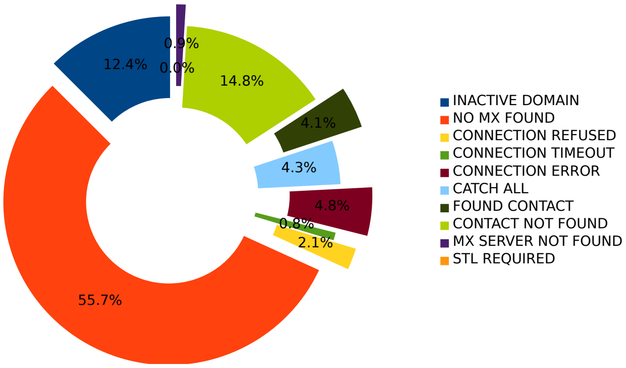

Figure 5: *.com* sample results

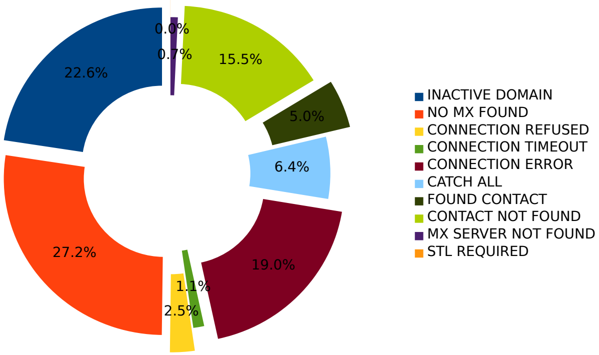

Figure 6: *.net* sample results

7 ANNEX