Analysis on MX-record queries of non-existent domains

Security and Network Engineering master
Research Project 2
July 5, 2020

Jasper Hupkens
University of Amsterdam
jasper.hupkens@os3.nl

Siebe Hodzelmans
University of Amsterdam
siebe.hodzelmans@os3.nl

Supervisor:
Jelte Jansen
SIDN
jelte.jansen@sidn.nl

Supervisor:
Cees de Laat
University of Amsterdam
c.t.a.m.delaat@uva.nl

***Abstract*—E-mail being sent to non-existing domains is a problem affecting every TLD in the world. In this research we propose a classification model for the organisation managing the .nl TLD, SIDN. We started with analysing a data set of queried non-existent domains, specifically where an MX record was queried. Non-existent domains are either domains which have previously expired or domains that contain typos made by people, meaning (in our case) they did not exist. Several of these domains were registered by us and the e-mail on these domains was collected. After some of the domains were found to be receiving sensitive information, we looked at data surrounding these domains. With this information we have created a classification model based on the Levenshtein distance to existing domains which are likely to receive sensitive information. We have provided this model to SIDN.**

Keywords - DNS, expired domains, typo domains, MX queries, personal data, e-mail

I. Introduction

On the 10th of April 2019, a large data leak was reported at an institution in The Netherlands that protects the well-being of children, called Samen Veilig Midden Nederland (SAVE) [1]. This was done by a journalist from RTL news, who, in turn, was informed by whistleblowers who found the data leak.

This data leak occurred because an automated system was still sending e-mails to e-mail addresses of an expired domain: BJZutrecht.nl. When the whistleblowers registered the expired domain and set up a catch-all e-mail server[1], they witnessed e-mails coming in for the domain. According to the article [1], 3.278 dossiers were leaked involving 2.702 children.

This was not the first time a case like this presented itself. On the 20th of January 2017, Computable wrote a story on a similar case, this time involving the Dutch National Police [2]. In this case the police district had changed their name and let the old domain name expire.

That this issue is not only a problem restricted to the Netherlands can be seen in the research done by Szathmari [3], who researched the impact of expired domain names in the Australian Law sector. From these examples we can see that this problem lives across multiple sectors and multiple Top Level Domains (TLDs).

Because of the way the Domain Name System (DNS) is designed, only one organisation can give us access to the complete data set of which non-existent .nl domains are still receiving queries. This is the maintainer of the .nl TLD, Stichting Internet Domeinregistratie Nederland (SIDN) [4].

We have found SIDN willing to support us in this research by providing us with the necessary data. This data consists of the queries matching our desired Query Type (QTYPE), MX, and Response Code (RCODE), NXDOMAIN, from May 11th to May 17th 2020. With this data we investigate whether it is possible to classify non-existent .nl domains as having a high potential of receiving sensitive e-mail using solely the name of an non-existent domain name and the knowledge that it

[1]A catch-all e-mail server will accept all the e-mail for a given domain no matter which local-part it is addressed to.

has been queried for the reason of sending e-mail to it. To do this, we will register a number of domains to determine which domains receive sensitive e-mail. The intended result of our research is to provide SIDN (and potentially other TLD administrators and administrators of intermediate domains, e.g. *.co.uk*) with a method to classify non-existent domains receiving sensitive e-mail and thereby hopefully preventing future data breaches.

This paper is structured as follows: Section II outlines the research question and supporting sub-questions, Section III presents previous research done on non-existent/expired domains as well as giving a short theoretical explanation on concepts and technologies used. Section IV explains the research setup, whereas Section V focuses on describing the research methodology. Next, Section VI presents the results, followed by the discussion in Section VII, which discusses the results and their future application. Afterwards is Section VIII, the conclusion. We conclude with Section IX in which we present future work.

II. Research question

The main research question for this project is defined as follows:

> Is it possible to classify non-existent .nl domains, using Open Source Intelligence (OSINT), as having a high potential for receiving e-mail with sensitive content?

To support this research question the following sub-questions have been defined:

- What OSINT sources can be used for classifying domain names?
- What classifiers can be identified for a domain being the recipient of sensitive information?
- What classification system can be used for classifying domain names?

III. Related work and background

A. Related Work

As mentioned in Section I, similar research focusing on the Australian law sector has been done by Szathmari [3]. In this research the researcher showed that expired domains of law firms were still receiving numerous e-mails. During this research the researcher was able to:

- access confidential documents of former clients;
- access confidential documents of the former practice;
- access confidential e-mail correspondence; and
- access personal information of former clients.

However, the researcher in this case accessed all the received e-mail to see what the nature of the content was. As will be explained in Section V, our approach will be different. We want to start with less intrusive ways of determining whether an e-mail contains sensitive information.

Next, the TIDE project is conducted by researchers of the University of Twente [5]. Here, researchers worked on the detection of bad actors, for example spammers, by doing active measurements on domains using DNS, e.g. the number of MX records used by domains. These could indicate malicious use of a domain.

SIDN also provides a service called the Domeinnaambewakingsservice (DBS) [6]. In this service, owners of a domain name can enter a keyword representative of their domain name. This service will then detect several types of abuse surrounding this keyword, e.g. the detection of typo-squat domains being registered.

Lastly, Schlamp et al. researched the scenario of attackers being able to take over IP prefixes and AS numbers because they are registered on domains which are now expired [7]. They found that at the time of writing “73 /24 IP prefixes and 7 ASes are vulnerable to be stealthily abused”.

B. Background

1) Domain Name System

DNS was first described in RFC 882 by Mockapetris in 1983 [8]. It solved the need of a way to describe systems such that applications could reach these systems spanning “multiple administrative boundaries”.

DNS works with so-called Resource Records (RR). These records contain the data a host can request. One type of record is the Mail Exchange (MX) record.

Without taking caching into account, every e-mail which is sent, requires a DNS lookup for an MX record to happen first.

2) MX records

MX records are heavily used on the Internet. Before an e-mail can be sent, a Mail Transfer Agent (MTA) needs to know where to send it to. This is done by sending a DNS query with Query Type (QTYPE) 15. The information is collected from the MX records which describes which server handles the e-mail for that domain.

3) NXDOMAIN

Domain names consists of several parts, called labels. Their format is a consequence of the hierarchical nature of DNS. The domain name `example.com.` consists of three labels, separated by a dot. From left to right the labels are: `example`, `com` and “the empty label”. The empty label is used to point towards the root servers [8]. Non-existent domains occur when any of the labels following the empty label - which always exist - do not exist. In this research, when we refer to non-existent domains, we mean that it is currently not within the .nl zone file. If a domain does not exist within the .nl zone this is sent back to the resolver querying the domain in question. The Response code (RCODE) used for NXDOMAIN is 3.

Within this research we also use the terms expired and typographical error (typo) domains. Expired domains are domains that were once registered, but are now not registered anymore (i.e. they do not exist within the .nl zone). Typo domains are domains which look like a registered domain and could have existed before, but do not necessarily have had to. Nevertheless, both expired and typo domains are types of non-existent domains.

4) OSINT

Open Source Intelligence (OSINT) is defined as "the collection, processing, analysis, production, classification, and dissemination of information derived from sources and by means openly available to and legally accessible and employable by the public" by Schaurer and Störger [9]. In short, it is the use of information which is publicly available. For our research, the most notable examples are the NLRetro tool [10] and the Wayback Machine [11]. The specific use of these tools will be discussed in Section V.

IV. Research Setup

During this research we need a few components to be able to both process the received data and receive the e-mail after we have registered the selected domains.

For the first part we create an Elastic stack [12]. This is to process the data with visual representations. For example, how queries were spread out over time or where ASes are registered from which queries originated. We will be ingesting the raw data (the CSV files from SIDN, data about ASes, and the keyword lists) with Logstash in an Elasticsearch cluster and visualize it using Kibana. Using these tools we have the ability to enrich the data and search the data interactively with various combinations of filters.

Besides Elastic, part of the data processing will happen using command line tools such as $grep$, cut, wc and $sort$. We use these tools for basic processing and aggregation.

For the second part we have configured an MTA using Postfix [13]. However, just an MTA was not good enough since we had the following requirements:

- Save every e-mail to a different file (i.e. one e-mail per file) so maximum traceability can be ensured later on if we have to open and read specific e-mails to classify their content. This also enables us to delete specific e-mails without having to read them.
- Ability to send automatic replies on incoming e-mails, notifying senders that their e-mail is now part of our research.
- Being able to handle multiple domains and save these e-mails separated from other domains.
- Allow any local-part (the part in front of the @ sign) to be used, also known as a catch-all server.

This leads to the conclusion that these are requirements Postfix cannot fulfil all by itself. So to save the e-mails on a per file basis and to also separate these e-mails per domain we use Dovecot [14]. Dovecot is known as a Mail Delivery Agent (MDA). With some configuring we manage to get Dovecot to show the desired behavior.

The next requirement, auto reply on incoming e-mail, is also not a Postfix feature. Therefore, we will create this functionality ourselves using a bash script running as a cron job every minute.

With some configuration, we will make Postfix into a catch-all e-mailserver, fulfilling the final requirement.

V. Methodology

To be able to answer the research question and sub-questions, we have executed this research in several stages.

A. Collecting domains

We have received all DNS queries for MX records that SIDN has received in a week, but which were answered with NXDOMAIN. This suggests that e-mail was intended for this domain, and that the domain did not exist, or was quarantined at the time.

To protect the privacy of the users doing those lookups, SIDN hashed the resolver IPs with a unique salt per IP and these salts were not made available to us. This method still allowed us to distinguish unique resolvers without being able to trace the request to a specific IP address.

Taking a first look at the data we realized that a first selection had to be made. The data we received are the queries from May 11 till May 17, 2020. The data set consisted of millions of queries, many more than we could examine within this research. Therefore, we focused only on relevant domains. Relevant domains within this research are domains which are likely to receive personal data of a special category or other kinds of sensitive data. The special categories of personal data are defined in the General Data Protection Regulation (GDPR) [15]. It is likely that the model we are going to develop will use this first kind of filtering.

B. Selecting domains

We wanted an objective method to choose domain names to register, based on the data set provided by SIDN and other (OSINT) sources. These were the NL Retro tool, the Wayback Machine, and lists of keywords. The keywords were from several sectors (e.g law, financial, medical) and were found as a list per sector using Wiktionary [16]. When presented with equal candidate domains, we attempted to choose those that made our selection more diverse (i.e. more types of businesses). Besides the domains we selected, we also registered several domains picked randomly from the top 100 queried domains as a control group.

In the following stage, we registered the chosen domain names and set up a catch all e-mail server. We configured

this e-mail server with a spam filter to try and filter out spam messages. The e-mail that did come through the filter was most-likely legitimate and, because of the way the domain names were selected, potentially sensitive. We had to check whether that was indeed the case, to gather input data for the classification model. As these e-mails are not intended for us, we tried to minimize our infringement with the following approach:

1) After receiving an e-mail, we automatically send an e-mail to the sender, or the postmaster of the domain of the sender (if noreply@ is used), informing them that their e-mail had been received by us and the purpose of our research. We also requested that they informed us what they wanted us to do with the e-mail, and (if they were willing) to help us with our research by telling us what the nature of their e-mail was. We also specifically ask if the e-mail contained sensitive information. This e-mail is shown in Appendix A.
2) If after two working days we did not receive a reply on our e-mail we looked at the metadata of the e-mail, such as the subject, the names and file types of possible attachments and headers of the e-mail (i.e. anything but the content). Using this information we wanted to try to determine if the e-mail contains sensitive information.
3) If the previous step still left us with the suspicion the e-mail contained sensitive information, but we were not able to prove it we sought approval from our supervisor. We produced a short motivation as of why we wanted to read the e-mail. If the approval was given, we opened and read the e-mail. We did this as a last resort. After reading an e-mail (whether it was sensitive or not) we sent an e-mail to the sender's address informing we actually read their e-mail. At the end of this research, we informed the Ethical Committee which emails we have read.

After any step in the above process, if we determined that the domain received sensitive information, we marked the receiving domain as sensitive, removed it from the e-mail server configuration (so we did not receive e-mail for that domain anymore) and deleted all e-mails we received for that domain.

C. Creating a model

Based on the data we collected, we compared the domains which did receive sensitive e-mail with those that did not. Based on this data we will try to develop a model which can be used to classify other domains.

VI. Results

The results of this research are divided in five parts. First, we will discuss the analysis of the data we received which lead to the selection of the domains we registered afterwards. Afterwards we will talk about the gathering of the e-mail and how these were analysed. Next, the classification model is shown followed by our advice to SIDN on how to use this model. Finally, we will present an advice to the wider Internet community.

A. Data processing

To start this result section, we want to present the numbers from the data set we received. All data combined, we got 116.208.888 queries, covering 14.150.845 unique domains. These requests were made by 97.954 resolvers from 10.688 AS numbers.

To manually classify all these domains would be infeasible, therefore we needed to reduce the numbers.

Since the goal of this research is to verify if domains can be classified as having a potential for receiving sensitive information, we thought it might be best to try and filter on relevant sectors that potentially send and receive such information. In the end, we wanted to create a list with keywords which could be used to match the received queries against, to make a smaller, more relevant subset of the data.

We tried this in two ways, first using an in Section V described objective approach. However, when we started with the objective approach, we quickly came to the conclusion that there were still too many results for us to process. This was due to two reasons. First, several words in the Dutch language are very common parts of other words and therefore generate a lot of uninteresting matches. But even with those filtered out and only matching to words of six characters or more, we found there were too many domains to process. Domain names often consist of multiple words without any punctuation between the words, often combined to create another word. This "new" word means that more words match, also from different categories that have nothing to do with the domain names.

Therefore, the next approach was a subjective approach. Here, we made our own word list of words we knew. Either from manually looking through the data or from personal experience, which would be used in sensitive domains. The disadvantage of a subjective approach is that, if we will not think of a keyword, we do not get the results. Yet we found, as was to be expected with fewer keywords, that the number of results were also more manageable.

Eventually, we decided on a combination of both approaches. First we created a subjective list. Then, we went trough the results of the objective lists, adding words we found had interesting matches. We filtered all queries using this combined list. However, we noticed there were still a lot of false positives caused by certain keywords. These keywords were removed from the list to reduce the number of false positives. Next, all queries for sub domains were filtered out. We did this because we noticed that these, in the majority of cases, consisted of random strings. What they are used for is unknown. However, it is safe to assume these are not used to address e-mails to. We also discarded any query which

occurred less than 10 times on a single day. We found that whilst they would most-likely give the least results (given the number of queries), these queries formed the majority of the entries in the results. Finally, we filtered the results on whether they had a snapshot in the Wayback Machine or not to determine whether or not the domain was expired. Also, the NL retro tool of SIDN was used to determine whether a domain name had previously existed.

From this selection, we picked thirteen domains to be registered. We looked at the category of the domain and (e.g. lawyer, debt collector, dentist, general practitioner), the originating ASes of the queries, and at which point during the week the observed queries were made. Our aim with this analysis was to have a good variety of domains with a high chance of regularly receiving sensitive mails.

To also have a control group, we registered thirteen random domains from the top 100 queried domains[2]. Furthermore, we selected four typo domains which matched on our "sensitive" keywords. Whilst not the intent of our research, during our analysis we noticed a large amount of typos in domain names. We registered four typo domains as other possible sources of sensitive mails. From the thirteen random domains we chose, five could also be considered typo domains and therefore for this research we chose to classify these as typo domains. In the end this gave us thirteen expired domains, nine typo domains, and eight domains in the control group. This brought the total number of domains we registered to 30. These were registered at tim427.net who also provided the DNS infrastructure for us. A masked list of the domains can be found in Appendix B.

B. E-mail collection and processing

As mentioned in Section IV we set up an e-mail server based on Postfix and Dovecot. After the registration of the domain names was completed, the e-mails started to arrive. We collected e-mails for 14 days.

We gave everybody we received e-mail from, two work days to respond to our auto reply asking them if personal data of a special category was sent in the e-mail. In most cases however, no reply was received from the senders. Every day, we selected all the e-mails older than two work days and both the sending address and subject of the e-mail were combined in a file. E-mails marked as spam by Spamassassin were removed from this file. We reviewed every entry manually, to see whether it could either already be classified as being sensitive or not sensitive, or whether further inspection of the headers was needed.

In case further header inspection did not give the desired results, i.e. an e-mail could not be classified as being either sensitive or not, we marked the e-mail and explained the situation to our supervisor to get permission to read the e-mail. This permission was then given or not. All these decisions are stored in an overview and will be shared with the Ethical Committee after the project is completed, as was agreed prior to starting with this research.

If a domain was found to be receiving sensitive information, the statistics of how much spam (as determined by Spamassassin) were made. Afterwards, all e-mail for that domain was removed and the domain was deactivated in the configuration of our e-mail server. This prevented the server from accepting new e-mail for this domain. If an e-mail server reached out to our e-mail server for a deactivated domain, it replied to the server with: Relay access denied.

In the end, sensitive e-mail was found on 6 of the 30 domains registered. The domains which were found to receive sensitive information were (by their number): 10, 13, 22, 25, 26 and 28. One of these domains was an expired domain which belonged to a law practice. The others were all typo domains, which were similar to either Internet/e-mail service providers, or similar to domains belonging to healthcare organisations.

For some domains we were surprised to see that we did not receive any e-mail (except the one test e-mail we sent). Therefore, we compared the number of queries SIDN received for our domains, during the period we had our mail server active, with the number of mails we received. See Appendix B. As can be seen, for example with domains 15 and 21, SIDN received more then 13.000 queries, but we only received our own test email on those domains.

C. Classification model

Most of the sensitive e-mail we received was on typo domains. Therefore, we propose a fairly simple classification model. First, extract all domains from the queries which exceed more then 10 queries per day. The purpose of this threshold is to filter out queries from scanners and DNS measurement projects. Second, create a list of existing domains which likely receive sensitive information. From our research, we identified two categories: domain names of Internet/e-mail service providers and domain names of medical organizations. A domain name can be checked by computing the Levenshtein distance (a measure of how different two strings are) between the domain name and the domain names in the list, e.g. the Levenshtein distance between domain and domein is 1. If the lowest found Levenshtein distance is one (meaning only one character is different between the domain and a domain from the list) it is likely to receive sensitive information.

D. Advice to SIDN

Our advice to SIDN (and by extent other administrators of TLDs) is twofold. First, monitor, or even restrict

[2]These were filtered to remove all the domains which could not be registered and any domains with a sub domain, this left 83 domains from the top 100.

registration of typo domains. For example by requiring additional verification, proving a person/organisation has a link to the domain name before one is allowed to register such a domain name. The typo domains looking like those of an e-mail provider received a low percentage of e-mails containing personal data in one of the special categories. However, we also observed hundreds of mails revealing other personal data, e.g. several mails asking for account confirmation or to confirm password resets. These accounts were from financial institutions, social media accounts, telecom providers, etc. This data can easily be used by malicious parties for criminal purposes, such as identity theft and fraud. The other typo domains, which look like domains belonging to healthcare providers, had a very high percentage of e-mails containing personal medical data, which is one of the special categories of personal data. Sometimes just somebody's name and the fact they are communicating with a certain health institution is sensitive in itself.

Another suggestion we have is to monitor domains that are in quarantine, but still receive MX queries. Above a certain threshold, SIDN could proactively warn the former owners of the domain, if possible, that people are probably still trying to send e-mail to them.

E. Advice to the wider Internet community

If end-to-end encryption of e-mail (e.g. PGP) was ubiquitous, the problem we investigated would not exist. From this research, we have also identified several recommendations we want to make to the wider Internet community:

- When validating a form, also verify that the domain the user has provided, actually exists. This would protect against users having typo domains as e-mail addresses. Actively helping to lower the risk of sending sensitive information to unauthorized recipients.
- When an e-mail address is associated with an account, check regularly and/or check when a new action has been taken (e.g. an order is placed after a long time) that the e-mail address is still correct.
- Everyone should be extra careful when sending or replying to an e-mail address when they have not contacted that address in a while.
- Only decommission domains after one has ensured that no legitimate e-mail is being received on those domains.
- When a domain is decommissioned, administrators should ensure that all mail clients are reconfigured (i.e. all e-mail clients should be prevented from sending e-mail using the decommissioned domains).

VII. Discussion

The initial data set we received had lookups to 14 million unique domain names. A lot of those unique items were queries which appeared to be brute force domain names (e.g. queries for $btklhc.nl$, $btklhd.nl$, $btklhe.nl$, et cetera). Even when we do not consider these, we still registered a small subset of the total amount of domain names. A larger sample set would have been desirable. We made our choice based on cost of registering and the chance of finding personal data from a special category. Filtering domain names with keywords or domain list has another disadvantage. If organisations do not use one of the keywords on the list or use only their name, they will most likely be missed.

Our automatic reply did not always reach the intended destination. While we tried to account for noreply e-mail senders by sending the reply to $postmaster@domain$ instead, this sometimes failed. Either because the e-mail server did not accept e-mail on postmaster (against the Simple Mail Transfer Protocol specification [17]) or because organizations did not use the term "noreply" in the from address. Another reason why some of our e-mail did not always reach the sender, was because the e-mail relay that we used was still on some spam blacklists, because of a recent spam incident.

Currently our proposed classification model does not classify expired domains. Since we only had one expired domain that received sensitive information, we were not able to compare domain names and find characteristics which could be used in a classification model. In the end, we received fewer e-mails than we initially expected. This impacted our research since we had less data to work with. This was surprising given the fact that in the original data set, some domains received around 2000 queries in a week (also the case during the time we had these domains registered and were actively receiving e-mail). However, we checked the number of DNS queries received for these domains while we had the MTA running. During this period, these same domains received a similar amount of queries as before. In other words, despite thousands of queries, we received no e-mail for some domains. This can be seen in Appendix B.

VIII. Conclusion

In order to answer our main research question, we first need to elaborate on the sub-questions.

The first sub-question focused on whether domain names can be classified using OSINT. We found that this is partly possible. By using the NL Retro tool from SIDN, one can find out whether a domain name previously existed or not. Next, by using the Wayback Machine also a snapshot of the previous web page can be obtained. Although, occurring multiple times in the past, we have not been able to irrefutably prove that expired domain names receive sensitive information. Within this research we only received sensitive information on a single domain which was previously registered. Hence, snapshots are currently irrelevant, as a lot of typo domains have never been registered. If in the future it can be proven that sensitive data is received on expired

domains, the Wayback Machine could be a valuable service to use.

The second sub-question was: what classifiers can be identified for a domain as being the recipient of sensitive information? This sub-question was answered in multiple steps. First of all we found that given the data set, a start has to be made using a biased keyword list. This biased list is also the first part of the classification. We also found that domain names are often quite descriptive of the organisations they belong to. When these descriptive parts of organisations dealing with sensitive information are identified, they could be used as keywords. If domain names match to this list of keywords, they can already receive a first classification. Also we have comprised a list of domain names belonging to both healthcare providers and Internet/e-mail providers. This list can also be used for matching purposes. Resemblance to one of such domains will indicate a higher chance of receiving sensitive e-mail. The resemblance should have a maximum Levenshtein distance of one.

The third sub-question was about how the previously identified classifiers can be used in an automated way. We automated the use of the NL Retro tool from SIDN. This way the domains which are matched against one of the domains from the list are checked on being an expired domain or not. If the output is empty, the lookup for a SOA record can be used to check whether the domain is currently registered. If both are negative, the domain never existed. We have provided a proof of concept with the functionality described above and have provided SIDN with the source code.

To answer our main question "Is it possible to classify non-existent .nl domains, using Open Source Intelligence (OSINT), as having a high potential for receiving e-mail with sensitive content?", we can conclude that it is possible to do this. We proved that typo domains are a real issue. From the nine typo domains we registered, five received sensitive information covering five out of the eight categories of personal data of a special category, as defined by the GDPR. However, none of these domains were registered before we did so. Thus, it was not possible to do a classification based on historical data in our case. What was left is the classification based on the domain name and whether these are within the range of looking like an organisation which processes sensitive information.

We have advised SIDN to look at our classification model and investigate whether it is possible to develop a new service, alerting organisations about typos people make for the .nl zone, based on the classification model we made.

IX. Future Work

In this research we found that analysing requests for MX records can give a good insight in typo- and expired domains. However, other widely used record types, like A and AAAA, were not taken into account in this research. We think that analysing the NXDOMAIN responses for these requests could give an even better insight.

As mentioned throughout the research, after analyzing the data received from SIDN, we found that domains with typos in them might be as big of a problem as expired domains. People make mistakes and will keep on making mistakes in the future. Therefore, we think further research is needed on typo domains with regards to e-mail specifically, because little research has been done on this topic. Furthermore, though typosquatting research has been done before for other services like webpages, we also would like to encourage more research into DNS queries on typo domains on all types of records, to aid in the prevention of typosquatting.

Whenever classifying content was necessary, because no reply was received from the original sender, this was a manual process. It would be very interesting to develop a system which could autonomously determine whether an e-mail, or any other file for that matter, contained personal data (from a special category). Research involving personal data could then be conducted even safer since people would not have to classify the contents manually. Thus, information inside emails would not have to be disclosed to any people anymore.

The earlier described discrepancy between the amount of queries sent versus the amount of e-mail received is interesting. For now, this means that the amount of MX queries observed cannot be used as a direct correlation to the amount of e-mail sent. Further research, and probably cooperation with third parties will be needed to see what is really behind the numbers. Until then, the numbers can only be used as an indication that a domain still receives queries.

X. Acknowledgement

We want to thank our supervisor Jelte Jansen of SIDN for his support during our research. We would also like to thank the other people of SIDN who have helped us with their feedback and ideas. Also, we want to thank tim427.net for providing the registration of the domains at a discount.

References

[1] Daniël Verlaan. *Groot datalek bij Jeugdzorg: dossiers duizenden kwetsbare kinderen gelekt*. URL: https://www.rtlnieuws.nl/tech/artikel/4672826/jeugdzorg-datalek-dossiers-kinderen-utrecht-email (visited on 08/03/2020).

[2] Pim van der Beek. *Politie lekt data via verlopen e-mail-domeinen*. URL: https://www.computable.nl/artikel/nieuws/security/5929279/250449/politie-lekt-data-via-verlopen-e-mail-domeinen.html (visited on 08/03/2020).

[3] Gabor Szathmari. *Hacking law firms with abandoned domain names*. URL: https://blog.gaborszathmari.me/hacking-law-firms-abandoned-domain-name-attack/ (visited on 08/03/2020).

[4] SIDN. *About SIDN*. URL: https://www.sidn.nl/en/theme/about-sidn (visited on 08/03/2020).

[5] Olivier van der Toorn. *TIDE project*. URL: https://www.tide-project.nl/ (visited on 08/03/2020).

[6] SIDN. *DBS*. URL: https://www.sidn.nl/en/product/dbs (visited on 08/03/2020).

[7] Johann Schlamp et al. "The abandoned side of the Internet: Hijacking Internet resources when domain names expire". In: *International Workshop on Traffic Monitoring and Analysis*. Springer. 2015, pp. 188–201.

[8] P. Mockapetris. *DOMAIN NAMES - CONCEPTS and FACILITIES*. URL: https://tools.ietf.org/html/rfc882 (visited on 08/03/2020).

[9] Florian Schaurer and Jan Störger. "The evolution of open source intelligence (OSINT)". In: *Comput Hum Behav* 19 (2013), pp. 53–56.

[10] SIDN. *NL-Retro (beta v 0.1) by SIDN Labs*. URL: https://nlretro.sidnlabs.nl/demo/ (visited on 30/06/2020).

[11] Internet Archive. *Wayback Machine*. URL: https://web.archive.org/ (visited on 08/06/2020).

[12] Elasticsearch. *Elastic Stack and Product Documentation*. URL: https://www.elastic.co/guide/index.html (visited on 07/06/2020).

[13] Postfix. *Postfix Documentation*. URL: http://www.postfix.org/documentation.html (visited on 07/06/2020).

[14] Dovecot Authors. *Dovecot manual*. URL: https://doc.dovecot.org/ (visited on 07/06/2020).

[15] Privazyplan. *Processing of special categories of personal data*. URL: https://www.privacy-regulation.eu/en/9.htm (visited on 02/06/2020).

[16] Wiktionary. *Categorie:Medisch in het Nederlands*. URL: https://nl.wiktionary.org/wiki/Categorie:Medisch_in_het_Nederlands (visited on 07/06/2020).

[17] J. Klensin. *Simple Mail Transfer Protocol*. URL: https://tools.ietf.org/html/rfc5321 (visited on 12/06/2020).

APPENDIX A
AUTO REPLY EMAIL

For English, please look below.

Beste meneer/mevrouw,

Wij zijn studenten aan de Universiteit van Amsterdam (Security and Network Engineering master, zie os3.nl) en voor ons afstudeeronderzoek doen wij onderzoek naar verlopen domeinnamen binnen het .nl Top Level Domain (TLD). Specifiek onderzoeken wij of op voorhand te bepalen is of een domein een hoge kans heeft om gevoelige e-mail te ontvangen. In het bijzonder gaat het hierbij om e-mail met bijzondere persoonsgegevens erin.

Hierbij een link naar de autoriteit persoonsgegevens over wat precies wordt verstaan onder bijzondere persoonsgegevens:

https://www.autoriteitpersoonsgegevens.nl/nl/onderwerpen/algemene-informatie-avg/mag-u-persoonsgegevens-verwerkenwat-verstaat-de-avg-onder-bijzondere-persoonsgegevens-6339

Wij willen u via deze e-mail graag informeren dat het domein waar u een mail naar stuurde tot kort geleden ook verlopen was. Voor ons onderzoek hebben wij dit domein weer geregistreerd en hebben daarmee uw e-mail (met het onderwerp zoals in de onderwerpregel beschreven staat) ontvangen. Dit hebben wij gedaan om onze hypothese te kunnen testen dat wij vermoeden dat er op het domein dat u benaderd heeft, inderdaad nog e-mail ontvangt met bijzondere persoonsgegevens. Naast domeinen waarvan we denken dat ze gevoelige informatie ontvangen, hebben we ook een aantal andere domeinen geregistreerd als controle groep.

Omdat het hier mogelijk om gevoelige gegevens gaat willen wij hier zorgvuldig mee omgaan en willen we voorkomen dat we de e-mail moeten openen om de inhoud te kunnen bepalen. Daarom zouden we u graag willen vragen om op deze email te reageren en daarbij de vraag te beantwoorden: ”Zaten er in de e-mail die u onlangs verstuurd heeft naar dit domein bijzondere persoonsgegevens en zo ja, om wat voor gegevens ging het?”.

Indien we antwoord van u ontvangen zullen we uw mail die u in eerst instantie had verstuurd altijd verwijderen en wij zullen ook geen gegevens van u opslaan. Indien u aangeeft dat uw e-mail inderdaad bijzondere persoonsgegevens bevatte zullen wij het domein markeren als gevoelig, verwijderen van onze server om geen nieuwe e-mails te ontvangen en alle tot op heden ontvangen e-mail verwijderen.

Mocht u vragen hebben dan horen wij die graag.

Met vriendelijke groet,

Siebe Hodzelmans en Jasper Hupkens

English:

Dear Sir/Madam,

We are students of the University of Amsterdam (Security and Network Engineering master’s programme, see os3.nl) and for our gradiation project we are researching expired domainnames within the .nl Top Level Domain (TLD). Specifically we are researching whether it is possible to determine upfront if an expired domain is likely to receive sensitive emails. For our research the are emails which contains personal data which fall in the special categories of personal data.

See the link below to see what kind of data falls in these special categories:

https://www.privacy-regulation.eu/en/9.htm

Through this way we would like to inform you that the domain you have tried to send a mail to has expired. For our research we have registered this domain and this made it possible to receive your email (the subject of that

email can be found in the subject of this email). We are doing this to make it possible to proof our hypothesis that the domain you tried to reach indeed still receives email which might contain personal information which falls in one of the special categories. Besides these domain which we suspect receive sensitive information, we have also registered a set of random domains which act as a control group.

Because it might be about sensitive information we want to treat it accordingly. Therefore we want to prevent the situation in which we will have to open and read the email in order to classify its contents. Therefore we would like to ask you the following question: ”Did the email you recently sent contain any data which falls in one of the special categories, and if yes, what kind of information was this?”.

In case we receive an answer from you we will always remove the original mail and no details about you will be saved whatsoever. In case you say your email indeed contained personal data within one of the special categories we will mark the domain as sensitive, remove the domain from our server to prevent receiving new emails and remove all the emails we already received.

In case you have any question please don’t hesitate to reach out.

Kind regards,

Siebe Hodzelmans and Jasper Hupkens

APPENDIX B
REGISTERED DOMAIN NAMES (MASKED)

Domain name number	Domain name related to:	Type of domain	MX queries	# of Mails	Levenshtein distance
1	Law practice	Expired domain	8	1	n/a
2	Healthcare provider	Expired domain	64	16	n/a
3	Law practice	Expired domain	0	1	n/a
4	Healthcare organisation	Expired domain	1906	1	n/a
5	Healthcare provider	Expired domain	25	4	n/a
6	Healthcare provider	Expired domain	177	12	n/a
7	Law practice	Expired domain	63	14	n/a
8	Healthcare provider	Expired domain	88	25	n/a
9	Healthcare organisation	Expired domain	196	17	n/a
10	Law practice	Expired domain	60	30	n/a
11	Debt collector	Expired domain	1982	1	n/a
12	Healthcare provider	Expired domain	4	12	n/a
13	Healthcare provider	Expired domain	206	2	n/a
14	Regular business	Control group	3	2	n/a
15	Regular business	Control group	13864	1	n/a
16	Regular business	Control group	13152	1	n/a
17	Regular business	Control group	13424	1	n/a
18	Regular business	Control group	13690	9	n/a
19	Regular business	Control group	14039	98	n/a
20	Regular business	Control group	13661	21	n/a
21	Regular business	Control group	13807	1	n/a
22	Internet service provider	Typo domain	38320	747	1
23	Internet service provider	Typo domain	23013	564	1
24	E-mail provider	Typo domain	422	134	1
25	Internet service provider	Typo domain	26623	573	1
26	E-mail provider	Typo domain	9727	1620	1
27	Healthcare provider	Typo domain	2	2	1
28	Healthcare provider	Typo domain	1041	50	1
29	Healthcare provider	Typo domain	1	1	1
30	Governmental organisation	Typo domain	14	1	3