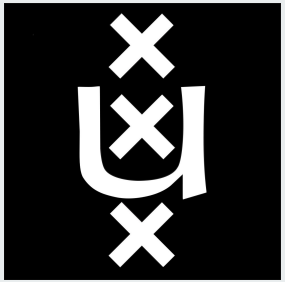
Analysis on MX-record queries of non-existent domains

Jasper Hupkens en Siebe Hodzelmans



Introduction

- Data breach at Samen Veilig Midden Nederland in April of 2019
- 3200+ files about 2700+ children were exposed because of an expired domain
- Issue proven multiple times covering multiple Top Level Domains (TLDs)
- In most cases sensitive information was retrieved

- This research was done with help from Stichting Internet Domeinregistratie Nederland (SIDN)

Research questions

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

Research sub questions

- 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?

Background information

- The Domain Name System (DNS) is primarily used for finding an IP address for a given domain name
- DNS is a hierarchical system where queries:
  - Start at the root,
  - Go to a Top Level Domain (TLD) (e.g. .nl),
  - Name server of an organisation.
- At the name server of an organisation so-called Resource Records (RRs) can be obtained.
  - A (QTYPE=1)
  - MX (QTYPE=15)
- If domain names are not in a zone of a name server, a name server will return Non-Existent Domain (NXDOMAIN, RCODE=3)

Background (2)

- Open Source INTelligence is information which is freely available on the internet
  - Google
  - Bellingcat
  - NL Retro
  - Wayback machine

Background information (3)

For this research, sensitive content was personal data which fits in one of “special categories” as defined in the General Data Protection Regulation (GDPR)[1]:

- racial or ethnic origin,
- political opinions,
- religious or philosophical beliefs,
- trade union membership,
- the processing of genetic data,
- biometric data for the purpose of uniquely identifying a natural person,
- data concerning health,
- data concerning a natural person's sex life or sexual orientation

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

Methodology

The following approach will be used within the research:

- Analyse and filter the data set,
- Select the domains,
- Receive the email,
- Classify content,
- Create classification model.

Protocol

Protocol for classifying contents of an email:

- Send auto reply
- Wait two days
- Inspect headers
- Ask permission from supervisor
- Give an overview to the Ethics Committee

Experiment setup

- An ELK cluster spanning three servers for data analysis
- A Mail Transfer Agent (Postfix)
  - Able to handle multiple domains
  - Catch-all (any local part (part before the ‘@’) can be used for all the domains)
  - Every mail is saved as a single file
  - Auto-reply to sender

The data

- All the DNS queries made to SIDN (.nl zone) for MX records, where the domains queried did not exist (i.e. NXDOMAIN was send as an answer), within the timespan of a week.
- May 11th to May 17th 2020

The data

- All the DNS queries made to SIDN (.nl zone) for MX records, where the domains queried did not exist (i.e. NXDOMAIN was send as an answer), within the timespan of a week.
- May 11th to May 17th 2020

116.208.888

The data (2)

- 14.150.845 unique domains
- 97.954 resolvers
- 10.688 Autonomous Systems

Filtering the data

- Basic Linux command line tools (grep, sort, cut, wc) and an ELK cluster
- Objective and a subjective keyword list
- Entries per day, if domain occurred less than 10 times it was removed
- Wayback Machine to see if they had exists before or not
- These were added as tags in ELK for further analysis

14.150.845

Filtering the data

- Basic Linux command line tools (grep, sort, cut, wc) and an ELK cluster
- Objective and a subjective keyword list
- Entries per day, if domain occurred less than 10 times it was removed
- Wayback Machine to see if they had exists before or not
- These were added as tags in ELK for further analysis

302.739

Filtering the data

- Basic Linux command line tools (grep, sort, cut, wc) and an ELK cluster
- Objective and a subjective keyword list
- Entries per day, if domain occurred less than 10 times it was removed
- Wayback Machine to see if they had exists before or not
- These were added as tags in ELK for further analysis

6.847

Filtering the data

- Basic Linux command line tools (grep, sort, cut, wc) and an ELK cluster
- Objective and a subjective keyword list
- Entries per day, if domain occurred less than 10 times it was removed
- Wayback Machine to see if they had exists before or not
- These were added as tags in ELK for further analysis

1.887

Domain selection

In total 30 domains were registered:

- 13 expired domains
- 13 control group domains
- 4 typo domains

Domain selection (2)

Because five of the randomly selected domains were typo domains as well, we ended up with the following split:

- 13 expired domains
- 8 control group domains
- 9 typo domains

Results

Overall

- 3966 emails received
- 35,1% spam (as classified by Spamassassin)
- 9 responses from natural persons
- 5 emails read
- 5 out of 8 special categories of personal data encountered
- 5 typo domains, 1 expired domain marked as sensitive

Special personal data encountered

In the end we encountered the following special categories of personal data in the emails:

- political opinions
- religious or philosophical beliefs
- trade union membership
- data concerning health
- data concerning a natural person's sex life or sexual orientation

Other sensitive data encountered

Other sensitive data encountered in the emails:

- Multiple accounts (Financial institutes, Social media, Telecom, etc.)
- Curriculum Vitaes
- Credit card verification
- A GPS tracker report
- Data addressed to a law firm

The unexpected ones

- Hardly any email on the expired domains, though initial data said several hundreds/thousands of queries per week
- A lot of (sensitive) mail on typo domains
- After 3 days we already received personal data in multiple of the special categories on 4 domains

Typo domains

- After analysis of the mail log of the server we found 1600 local parts (the part in front of the ‘@’) used to send email to our typo domains
- Therefore we can conclude that it was not e.g. a single person generating all the queries and this is indeed a widespread problem

Classification model

- Domain is queried > 10 times a day
- Create a list of existing domains likely to receive sensitive information
  - mail/internet providers
  - medical institutions
- Calculate Levenshtein distance of domain compared to list
- Distance < 2: likely receiving sensitive email

Conclusion

- Classification using OSINT is possible
- Main classifier we found is Levenshtein distance

“Is it possible to classify non-exitent .nl domains, using Open Source Intelligence (OSINT), as having a high potential for receiving email with sensitive content?”

Yes, but we have only been able to prove this for typo domains

Discussion

- More domains, longer time
- Subjective word list
- Large discrepancy between queries and actual email
- Limited proof for a problem regarding expired domains

Future research

- Typo domain research, also for other record types (A, AAAA, etc.)
- Create classification model also for other TLDs
- Automated system for recognising personal data in email

Final remark

Where possible, the typo domains were handed over to the organisations to which the original domain belonged to.

Questions?

Key take-aways:

- Do not let domains expire until you are sure no (legitimate) mail comes in anymore
- Typo domains are just as big of an issue as expired domains
- These problems will probably exist in other TLDs and hopefully research will start there as well