LEMMINGS: Preventing data leaks at TLD scale (Tech Report)

Maarten Wullink
SIDN Labs
Arnhem, The Netherlands
maarten.wullink@sidn.nl

Moritz Müller
SIDN Labs and University of Twente
Arnhem, The Netherlands
moritz.muller@sidn.nl

Giovane Moura
SIDN Labs and TU Delft
Arnhem, The Netherlands
giovane.moura@sidn.nl

***Abstract*—Email is one of the most important means of communication on the Internet. On a daily basis, countless emails are sent and received, of which some may contain sensitive information. Domain names are used to identify senders and recipients, are usually leased for a period of time, and thus, can change ownership. The latter has led to data leaks in the past. For example, a Dutch journalist registered a number of domain names formerly owned by the national police and was able to receive and read sensitive emails that were still being addressed to the previously defunct email addresses.**

Our goal is to take steps into raising awareness and preventing such data leaks on a large scale. In this paper, we propose a privacy preserving approach that relies on the fact that the Domain Name System (DNS) plays a crucial role in email communication and on the unique role of registries of Top-Level-Domain (TLD) names like `.com` or `.net`. Our approach uses DNS traffic to identify attempts to send email to addresses linked to deleted domain names before the domain name becomes available for everyone to register. We implement our approach in a tool named LEMMINGS and deploy it at a top-10 country-code TLD. We discuss challenges of this approach and measure its impact. LEMMINGS is now running continuously for more than eight months, has warned over 54 thousand domain name owners about potential data leaks, and has received positive feedback from domain name industry stakeholders.

I. Introduction

Domain names like `example.com.` play an important role in the communication of organizations and individuals worldwide. They are the front door to websites and enable email communication, which, despite the introduction of other messaging tools like Slack or Discord, is still growing [23].

The domain namespace is structured in a tree-like way, with the root (.) on top, followed by the so-called Top Level Domains (TLDs), such as `.com` and `.tokyo`. Individual *registries* manage the TLDs and are free to set their own procedures and polices. Users (so-called *registrants*) may register domain names with their TLDs of choice via a so-called *registrar* (or resellers). In case of `.AnonCc` more than 1,000 registrars and more than 46k resellers are responsible for registering domain names on behalf of the registrants.

Usually, registries lease domain names for a period of time instead of selling them perpetually. In this way, if a domain is no longer in use, it can be reclaimed by another future user. Domain names are typically leased for a year and if the domain is not renewed they *expire* and are returned to the namespace. Depending on the TLD policy, anyone could register the expired domain after that. Figure 1 shows the evolution of the namespace of the `.AnonCc`, the ccTLD of *Anonymized EU Country*, highlighting that a large share of domains expires daily.

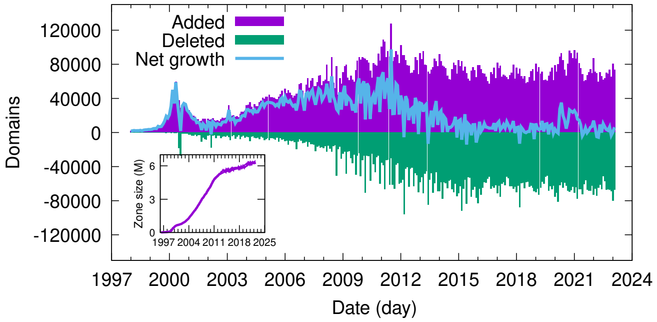

Fig. 1. Daily changes in the zone: added domains, deleted domains, and total number of domains between 1997 and 2023.

While registrants can delete domain names, it does not mean that other users will stop trying to reach them, e.g. because they are unaware that the domain name is not active anymore. This effect is known as “residual trust” [12] and has been used by the drop-catch domain industry [10]. Those are individuals or organizations that re-register deleted domain names in order to place ads on them or sell them [9].

Another reason to re-registered deleted domains is more nefarious: someone may register a deleted domain name in order to impersonate the previous registrant or *silently* collect traffic directed to the domain. This is particularly worrisome for asynchronous services as email, in which the new registrant could set up email servers to collect traffic to a former deleted domain, and, in this way, collect all incoming email that unaware users may send.

Such attack has been deployed and made the news in the Netherlands, where an investigative journalist was able to collect more than 3,000 reports on more than 2,700 children, containing sensitive data on drug abuse, family issues, psychology issues and sexual violence [6]. The data was obtained by registering a deleted domain that previously belonged to a child services institution and setting up an email server on the domain. By collecting emails that were automatically sent by a reporting system to the mail servers of the domain, the journalist retrieved this data. A similar incident occurred when a journalist registered domain names formerly used by the Dutch police [3].

Both examples demonstrate that a well-motivated attacker could incur serious damage. However, it remains an open question how to prevent such attacks. Holding on to not longer needed domain names just to prevent such attacks places the costs on the registrant side but may be feasible depending on the domain name. Also, registries often put expired domain names under quarantine, meaning for a period of time only the registrant can reclaim it. Yet, this quarantine period can only delay future attacks but does not prevent them. Moreover, many registrants may not even be aware of the risks of such attacks – they may simply lack the technical skills.

In this paper, we leverage our position as a ccTLD registry to try to prevent such attacks. We attempt to achieve this goal by identifying potentially vulnerable domain names after they have been deleted but before they left quarantine. Then, we provide their registrants with a final notification, via e-mail, where we explain the risks associated with the domains being re-registered by a third party. With this notification we want to make registrants aware of such risks and let them decide for themselves if it is worth to retain the domain, whether they would like to inform their peers about the deleted domain name, or whether they would not like to take any actions. By deploying such an approach at a registry we can protect all domain names registered under a TLD at once. This is more effective compared to when every individual registrar and reseller needs to take action. Ultimately, our hope is to *prevent* data leakage through silent e-mail harvesting and to inspire other registries to follow our approach.

To that end, we make the following contributions. First, we present an approach to determine which domain names might be vulnerable to such attacks (Section III), based on their Domain Name System (DNS) traffic observed at the authoritative DNS servers of the `.AnonCc` zone and domain name attributes. Then, we describe how we implement this approach in a tool called deLetEd doMain MaIl warNinG System (LEMMINGS) (Section IV). We then deploy it for a 8.5 month period, evaluating more than 587 thousand pending delete domain names in the `.AnonCc` zone, notifying the registrants of 54 thousand (Section V). We show that even though only a small fractions of the notified registrants decided to re-register the domain, our warnings are overall perceived useful and have, according to some registrants, prevented potential data leaks.

II. Background

Our approach mainly relies on data collected at the domain registration system and in the DNS and we explain both in more detail in this section.

A. Domain Registration Life-Cycle

We exclusively analyzed domain names of the country-code Top Level Domain (ccTLD) for `.AnonCc`. Therefore, the domain life-cycle described here is valid in the context of `.AnonCc` ccTLD, although other ccTLDs may have similar life cycles.

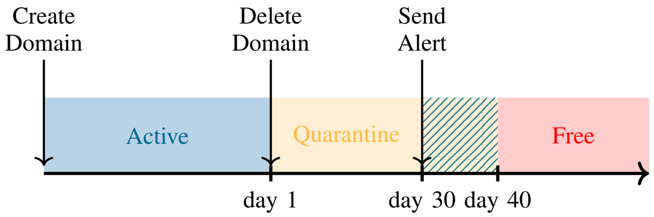

Fig. 2. Simplified domain life cycle (if not renewed).

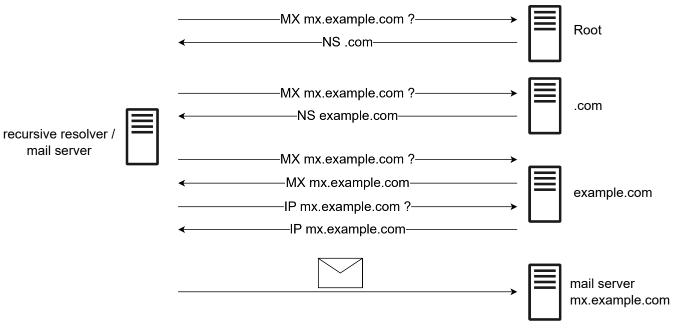

Fig. 3. DNS queries sent sequentially by a resolver before sending a mail to the mail server of example.com with an empty cache. For simplicity, the resolver also acts as a mail server.

The domain life cycle for these domains consists out of multiple states and state transitions. Figure 2 shows a simplified version, only the relevant states and events are shown. When a registrant registers a new domain, the domain enters the “active” state. In this state the domain is published in the `.AnonCc` zone if a valid name server set is also received from the registrant. When the domain owner requests the domain to be deleted, then the domain will transition to the Redemption Grace Period (RGP) also known as “quarantine”. During this period, only the former registrant is able to restore the domain back to the active state. The domain remains in the quarantine state for a maximum of 40 days. After the 40 day quarantine period has been completed a domain can transition to the “free” state – it is then publicly available for a new registration. When the former registrant restores the domain before the quarantine period has passed (“cancel-delete”) then it transitions back to the active state.

B. DNS

The DNS plays an important role in the transport and delivery of email. If registrants would like to receive emails for addresses associated with their domain name (e.g. `foo@example.com.`) then they need to set up a mail server, publish the address of the mail server in the DNS (e.g. `mx.example.com.`) and add a MX-record to the zone of `example.com`. Thereby, they state that email addressed to `foo@example.com`. should be delivered at `mx.example.com`.

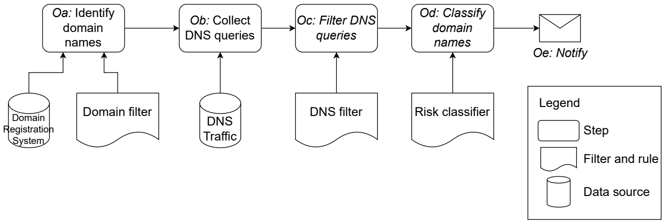

Fig. 4. Schematic flow of the proposed approach to prevent data leaks at TLDs. LEMMINGS implements this approach.

Before delivering emails, mail servers of the senders need to look up the IP address of `mx.example.com.` and Figure 3 shows the exchanged DNS messages. Here, they usually employ *recursive resolvers* that traverse the DNS hierarchy querying for the `MX` record of `example.com.` until they receive the answer or until they conclude that it cannot find the requested information. Every time a recursive resolver does not have any information of `example.com.` in its cache, it also sends a query to the authoritative name servers of the TLD (in this example `.com`).

During the quarantine period, the domain name `example.com.` is not published in the zone of `.com`. This also means that queries asking for the MX-record of the domain will receive a non-existent domain (NXDOMAIN) response and the email cannot be delivered. Note that DNS Query Name Minimisation [5] changes this behavior slightly and we discuss its impact on our approach in Section VI.

C. Mail

When a domain is deleted and its state has moved to quarantine or free, there are no longer any functioning email addresses linked to the domain. However, this does not necessarily refrain email senders to *attempt* to deliver mail. In that case, the mail server will not be able to deliver the message to the email recipient and the sender will receive a error reply back from the SMTP server stating that delivery has failed, also called a bounce. Only the sender receives this bounce but not the previous owner of the domain – the previous owner is unaware of the messages that are still being sent to the domain. When the email sender only sends infrequent mail to the deleted domain, it is also possible that the sender may never receive a bounce message. This might happen when the first email sent to a deleted domain is sent after the quarantine period for the domain has ended and the domain has already been re-registered with a catch-all email configuration. In cases where the sender is an automated system, the sender might not be able to process the bounce correctly and might not detect that the recipient address no longer exists.

III. Approach

Our approach relies on the fact that registries know when domain names are deleted and that they can observe DNS MX queries for these domain names at their authoritative name server infrastructure. Any observed MX query can indicate an attempt by a mail server to deliver email. If we can classify with a certain level of certainty that an MX query is the result of a mail server sending potentially sensitive information, then we want to warn the former registrant.

Our approach consists out of two phases: one preparation-phase before starting with notifying registrants and one processing-phase that runs continuously. In this section, we describe these phases on a high level and Figure 4 visualizes the general workflow. In the sections that follow, we describe how we implement each phase at `.AnonCc`.

A. Preparation

Before registries can start warning registrants, they need to do some preparation. Some of it is of technical nature, others focus on communication and coordination.

***P-a* Filter generation** Not every MX query observed at an authoritative name server results in an attempt to deliver an email relevant to the former registrant. For this reason, registries need to create filters discarding MX queries that might be the attempt to deliver SPAM, phishing, marketing emails, or other unsolicited emails or emails with low priority.

For instance, filters can be based on the origin of the MX query, and could include IP-address reputation, AS-type, or resolver query behaviour. In some cases, filters should be generated dynamically, e.g. when trying to filter out queries by resolvers that have not been observed before.

***P-b* Risk classification** Registries need to define which criteria need to be met before they inform the former registrant about the risk of a data leak. These criteria can include the number of relevant MX queries over a period of time, attributes of the domain name, and attributes of the registrant. Here, a registry also needs to strike the balance between warning as many of the registrants that are actually at risk and warning too many registrants unnecessarily.

***P-c* Notification** The type of message sent to the former registrant is crucial for achieving the ultimate goal. The message should convey the risk clearly and in a trustworthy manner, but without sounding alarming.

***P-d* Stakeholder coordination** In case of questions or complaints by the registrants, help desks by the registry and by registrars should be informed about this effort to handle calls appropriately.

In order to avoid interactions with other parties, our recommend approach is that registries send warnings to the registrants themselves. However, registries usually do not interact with registrants directly. Registrants register domain names at registrars, which also coordinate billing and extension of the registration. For this reason, another option is to allow registrars send the warnings on behalf of the registry.

B. Operation

After the registry has taken the preparation steps, it can start assessing domain names and sending out warnings. The DNS activity for a deleted domain names gets tracked for a period long enough to gain insights into relevant MX queries, but shorter than the quarantine period. For example, the registry could track a domain name for $3/4$ of the quarantine period to collect enough data but also to leave enough time for the registrant to take actions after receiving a warning.

***O-a* Identify relevant deleted domain names** Not every deleted domain name is suitable for being tracked. For example, because the registry does not have a functioning email address of the former registrant or because the registrant has never configured an email server and thus the risk of a data leak is already low. After this phase, only these domain names remain of which the registrants can likely be reached by the registry.

***O-b* Collect DNS queries** After applying the first filter, the MX queries to the remaining domain names are tracked. Here, the registry needs to be able see the source of a DNS query, the query type, and the domain name for which the resolver is asking.

***O-c* Filter DNS queries** In order to retain only those queries that might result in the attempt of sending a sensitive email, the registry now applies the filters generated previously in *P-a*.

***O-d* Classify domain names** The queries that remain after the previous step can give an indication to which extent a former registrant might run the risk of a data leak. By applying the risk classification rules defined in *P-b*, the registry can decide which domains need to be part of a notification.

***O-e* Notify registrant** Now, registries can notify the former registrants of domain names that the system has classified as running the risk of causing a data leak. In this approach, this means sending an email earlier composed in *P-c*.

IV. System

A. LEMMINGS Data-Sets

We implement the approach described in Section III at the `.AnonCc` ccTLD in a system called LEMMINGS. Before we describe the implementation of LEMMINGS in detail, we describe the used data-sets.

LEMMINGS uses multiple data sources. Newly deleted domain names are extracted from the Domain Name Registry System (DRS). DRS is the authoritative database for all domains registered in the `.AnonCc` zone. Additional information is retrieved from data collected by a custom developed web crawler, which crawls all registered domains once every month. Passive DNS data about DNS queries processed on the authoritative name servers of `.AnonCc` is extracted from the ENTRADA system [25]. Finally, LEMMINGS uses well-known sources of malicious domains such as the Spamhaus blocklist and APWG.

Domain Registration System The Domain Registration System (DRS) is the authoritative information source for all `.AnonCc` domains. DRS contains the domain related information attributes such as, but not limited to, the registration date, status, owner, technical contact and authoritative name servers.

Crawler Every month, we crawl the content found on websites linked to all active domains in `.AnonCc` with a self-developed web crawler. The crawler performs a shallow crawl and classifies the website based on the HTML content into a content category. The results are stored in a multi-year longitudinal data-set for the entire zone of `.AnonCc`.

Passive DNS We rely on ENTRADA [25], which is an open source passive DNS database system. Over 4 billion DNS queries are captured daily at the authoritative name servers for `.AnonCc`. These are forwarded to ENTRADA in the form of pcap files. ENTRADA converts the data from the pcap format to the Apache Parquet format, which is an efficient column-oriented data format – both in terms of storage requirements and analytical processing. The ENTRADA database currently contains 2,7 trillion records, requiring 400TB of storage, the data is stored and analysed on a Hadoop based cluster.

Abuse feeds In order to filter DNS queries we rely on feeds containing IP addresses suspected to be involved in phishing, SPAM, or other malicious activity. These feeds are provided by APWG and Spamhaus [2].

Sinkhole Finally, we operate a sinkhole service for 34 domain names, previously involved in botnet command-and-control activities. IP addresses that resolve one of these addresses or try to connect to our sinkhole are used to filter MX queries.

B. Preparation

1) P-a *Filter generation:* Query filters are used to remove DNS queries that have a high probability of not being part of a legitimate mail transaction or are linked to a larger mail campaign from a mail service provider. These filters may be static or dynamic. Static filter do not change very often, they are manually maintained lists, for example of AS numbers or IP-Addresses. Dynamic filter are generated each day, based on historical DNS query data. We base our filters on behavior of spammers as described in related work, our own measurements, and domain knowledge. Table I lists the filters.

Abuse feeds This filter uses external abuse feeds to create a filter based on the DNS resolver IP address. The following abuse feeds are used: APWG [2], Spamhaus ZEN [24]. We do not consider MX queries from IP addresses listed here.

ASN We create manually a static list of AS numbers of networks known to be used by mail service providers (MSP). An MSP is often used by enterprises for sending large mailings such as customer newsletters or other sales related mail. We expect that this type of email does not have the potential for creating privacy issues or data leaks, and DNS queries sent by IPs located in a network of a MSP are filtered.

Consistent resolver We expect resolvers used by mail servers to have a stable query pattern over the course of the week. We expect that bursts of MX queries over the course of the week can be linked to SPAM runs. For this reason, we filter IP address that show bursty query patterns.

Country A static list of countries of which we observe a disproportionally high number of abusive activity. We select these countries by analyzing the location of IP addresses listed on the Spamhaus ZEN blocklist. We add those countries to our filter of which we find a disproportionally high number of IPs on the ZEN blocklist.

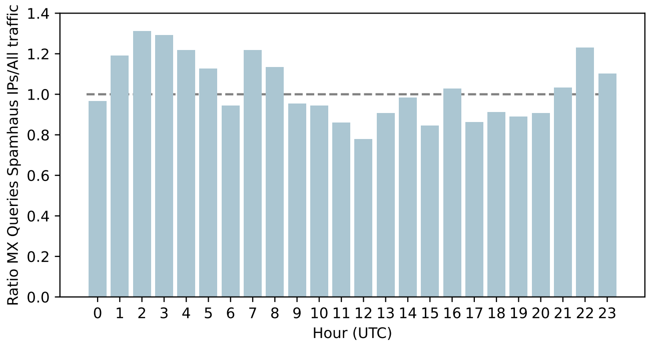

Fig. 5. Ratio MX queries from IPs on the Spamhaus ZEN blocklist and all MX queries observed over the course of one day.

High NXDOMAIN This filter removes DNS queries received from DNS resolvers sending an abnormal high rate of DNS queries for non-existing domain names. Through manual investigation of DNS traffic received from these resolvers, we learned that these are often used to systematically try to reverse engineer the zone of `.AnonCc` through the use of word lists causing a large number of responses with return code NXDOMAIN.

IP address We create manually a static list containing DNS resolver IP-addresses that have been flagged as abusive or are known to be linked to a Mail Service Provider.

New resolver This filter assumes that DNS resolvers who have only been seen first recently may have a lower trustworthiness. This is analogous to how some anti spam filter work [7], which assign a lower trust to newly registered domain names.

No Mail Not every domain name is used for mail. Using our crawler, this filter creates a list of domain names that do not have a mail server configured (no MX record present). We add DNS resolvers sending a high percentage of queries to domain names on this list to our filter.

Open resolver Open resolvers are resolvers that are open for a large part of the Internet to use, often due to a configuration errors. These resolvers are not to be mistaken with public resolvers such as Google DNS and Cloudflare DNS. Open resolvers are often misused by miscreants or show malicious behavior themselves [19] and therefore we assume that a large portion of the traffic from these DNS resolvers may be for malicious purposes.

Sinkhole `.AnonCc` operates a sinkhole where domain names linked to known botnets are hosted, the network traffic generated by the botnet clients is captured for analysis. This filter contains IP addresses of the botnet client.

Time Finally, we remove DNS queries that are sent within a specific time window. Figure 5 shows that IP addresses on the ZEN blocklist send disproportionately often MX queries between 1 am and 5 am UTC, compared to the overall MX traffic we see at `.AnonCc`. The time filter removes all DNS queries sent within this time window.

2) P-b Risk classification: The system uses a basic rule based classifier model to determine the risk category associated with each domain that needs to be alerted (see Algorithm 1). The model uses three distinct risk categories; Low, Medium and High to indicate the risk to the recipient of the alert. We decided to use three risk categories to communicate about the level of risk, because it might be difficult for the recipient to determine the risk based on a number of queries without having additional information.

A risk category is based on the number of received DNS queries, and other attributes linked to the domain, such as;

1) DNS query threshold
2) keyword match
3) SBI code
4) email address usage

DNS queries is the average number of daily DNS queries, received by a deleted domain in the first 30 days of the quarantine period, after removing unwanted DNS queries using the query filters. We have determined a lower threshold for each of the 3 risk categories based on the average number of DNS queries seen for all alerted domains - 1, 5 and 10 queries respectively. *keyword match* is true when the domain matches a list of keywords compiled for LEMMINGS. This list of keywords contains words linked to business activities known to make use of sensitive data, for example in the medical and legal field. The *SBI code* is a national code in *Anonymized EU Country* assigned to every business, used to indicate the type of business activity. It is based on the NACE code, the standard European nomenclature of productive economic activities. We use our web crawler to attempt to identify SBI codes on all websites using a `.AnonCc` domain. *email address usage* is true when our web crawler has identified email addresses, linked to the deleted domain, on any website using a `.AnonCc` domain in the last web crawl before the domain was deleted.

Algorithm 1 Risk classification algorithm

```
if $domain$ is keyword_match then
    $risk \leftarrow high$
else if $sbi\_code$ in sbi_high_risk_codes then
    $risk \leftarrow high$
else if $avg\_query \leq risk\_cat\_low\_max$ then
    if $mail\_address$ is used_on_web then
        $risk \leftarrow medium$
    else
        $risk \leftarrow low$
    end if
else if $avg\_query \leq risk\_cat\_medium\_max$ then
    $risk \leftarrow medium$
else
    $risk \leftarrow high$
end if
```

3) P-c Notification: We develop the text and design of the email sent to the former registrant together with communication experts, registrars, and a group of registrants. The email contains information about the deleted domain name, the risk category and the registrar at which the domain name was previously registered. Additionally, we provide information about the role of our registry, the potential risk associated with deleting a domain name, and recommend countermeasures to the registrant. These countermeasures include cancelling the deletion but also a recommendation to reach out to peers that have previously sent emails to the cancelled domain name to inform them about the fact that the domain name is getting deleted.

The email is composed in the language spoken in the country of the ccTLD, but also includes a link to a website explaining the risk in English. Furthermore, the email contains a link to a website containing answers to FAQ's.[1]

In case registrants delete multiple domain names on the same day and more than one domain name runs the risk of a data leak, then the email includes information to all affected domain names.

4) P-d *Stakeholder coordination:* Asides from the stakeholders mentioned above, we also involved our own support department before turning on LEMMINGS. We informed them about the goal of LEMMINGS and what questions they could expect from registrants.

C. Operation

Each morning, we collect all domain names that are on day 30 of their quarantine period and are thus released in 10 days (see Figure 2). Then, we proceed with the steps described below. At the final step, we notify the former registrants of domain names for which we have identified a risk for a data leak. In total, the process from creating dynamic filters until notifying registrants takes around 50 minutes.

1) O-a *Identify relevant deleted domain names:* Domain filters remove domain names unsuitable for monitoring with LEMMINGS. These filters look at attributes related to a domain names, such as the email address of the registrant.

No queries A filter for removing domain names for which no DNS queries have been processed in the 30 day period before deletion.

Age Domain names created and deleted within a relatively short period are probably not very valuable to the former registrant. Our assumption is that these domain names have not been used for sending important mail messages. Filtering young domain names also prevents the system from sending warnings to a registrant that maliciously registered a domain names for illegal activities, such as spam and phishing.

Privacy proxy Some registrants choose to register domain names using a privacy proxy service to hide their identity for different reasons. The chance that any alert message will be able to reach the actual registrant through the proxy service is low. The system will not send an alert to domain names registered though a privacy service. We identify privacy services through heuristics such as label matching in the domain name used for the mail address. For example, if the address contains labels such as "anonymous", "privacy" or "whois-protection" we assume that it is linked to a privacy proxy service. Additionally, we filter domains where the registrant email address is on a list of known privacy proxy service email addresses.

Unknown email In some cases we cannot find the email address in our registration database for historical reasons.

In-zone mail address Some registrants may use an email address linked to the same domain name as the deleted domain name. For example, a contact address of `info@example.com.` is linked to a deleted domain `example.com`. Because the domain name is not included in the zone when it is in quarantine, any mail sent to the domain name will result in a bounce.

2) *DNS analysis and notification:* The phases *O-b* to *O-e* are straightforward. We collect the DNS queries for the tracked domain names using ENTRADA and apply the filters developed in *P-a* to keep only relevant DNS queries. For domain names that are on day 30 of their quarantine period, we apply the risk classifier from phase *P-b*. Domain names that have at least a risk of *low* are notified automatically.

V. Results

The results in this section are based on data collected during an 8.5 month long period, starting early April 2022. During this period, we processed 587.778 domain delete requests and sent alert messages to registrants linked to 54.410 domains (9,2%).

Without applying these filters, we would not only have sent out emails for 10 times more domain names but we would also have warned 64% more individual registrants. Furthermore, 10 times more registrants would have received more than one warning per month. This shows that without the filters we would have ran the risk of our initiative being perceived as alarming, too broad and annoying and emphasizes the need for this more targeted approach.

Figure 6 shows how many domains have been deleted during each month of the data collection period. Also included in this figure is the number of alerts sent per month. During the first month no alerts have been sent, because deleted domains that have been added in that month, will not be alerted until day 30 of the quarantine period. The inverse applies to the last month, when fewer deleted domains have been added, because the observation period did not end on the last day of the last month.

In this section, we assess the effectiveness of the developed filters, the classification system, and finally whether LEMMINGS achieves its goal of preventing data leaks.

A. DNS Filters

The first set of filters have the goal to remove DNS queries that are probably not related to legitimate email transactions. The selection of these filters are described in Section IV-B1. During the 8.5 month period, LEMMINGS analyses over 106 million email related DNS queries for deleted domains. Our filters removed 75% of all received email related DNS queries, leaving 26,071,889 DNS queries that have probably been used for legitimate email transactions. Table I lists the share of

[1] Redacted for review

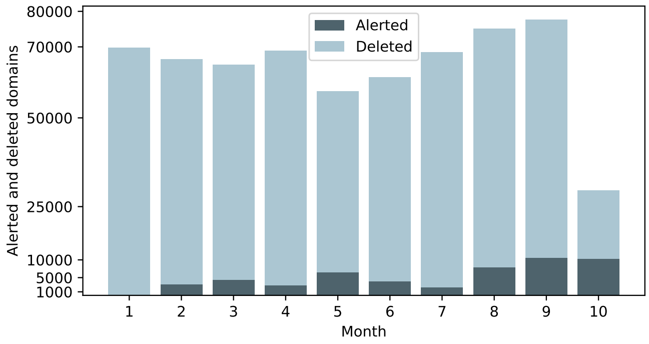

Fig. 6. Domain delete and alerts in period.

Name	Q removed	Share of queries
APWG (Abuse feed)	1,650	0%
Spamhaus (Abuse feed)	4,228,491	4.0%
ASN	47,177,603	44.4%
Consistent resolver	2,675,135	2.5%
Country	3,733,279	3.5%
High NXDOMAIN	38,125,287	35.7%
IP address	2,552,351	2.4%
New resolver	16,759,368	15.8%
No Mail	731,991	0.7%
Open Resolver	742,604	0.7%
Sinkhole	14,166	0%
Time	18,675,125	17.6%
Not filtered	26,071,889	

TABLE I
RESULTS PER FILTER.

queries to which each filter applied. One or more filters can apply to a query. Before filtering, the average number of mail related DNS queries per domain per day is 4.7. After filtering this drops to 1.2.

Especially the network-based (ASN), response-type based (NXDOMAIN), time-based, and the prominence-based (new resolver) filters are responsible for dropping large share of queries.

B. Risk Classification

After filtering DNS queries, we classify the risk for each domain name. Table II lists an overview of the number of alerts sent per risk category. The vast majority (77.85%) fall into the low risk category.

Risk cat	Alerted	%
Low	44,701	77.85%
Medium	8,080	14.07%
High	4,639	8.08%

TABLE II
ALERTS SENT PER RISK CATEGORY.

Name	Domains	%
High Risk SBI Code	7,089	12.5%
Keyword Match	17,534	31.0%
Email usage	22,538	39.8%

TABLE III
RESULTS PER CLASSIFICATION PARAMETER.

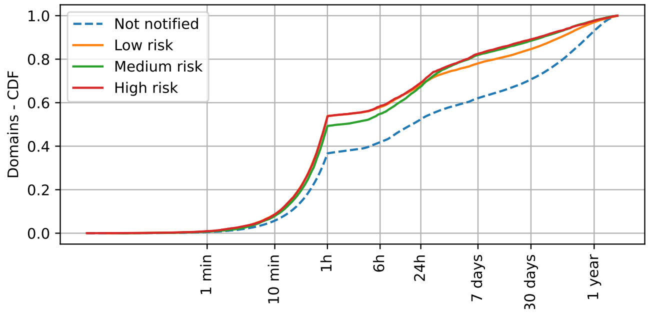

Fig. 7. Time between release out of quarantine and re-registration by risk category.

1) Non-DNS based classifier: Besides DNS queries, the classification algorithm uses three additional features, listed in Table III. Here we find that both the domain matching a keyword (31%) and the presence of a linked email address (39.8%) on a crawled `.AnonCc` domain website have a large influence on the outcome of the algorithm.

2) Domains at risk: The question is, however, whether these domain names were actually at a higher risk of having a data leak. This assessment is for us impossible to make directly. For this reason, we defer to the likelihood that domain names are re-registered after being released from quarantine.

After a domain name is released from quarantine it becomes available for general registration. We show that domain names that we classified as having a higher risk of a data leak have also a higher chance of being re-registered. 16% of domain names not warned by LEMMINGS were re-registered. In contrast, 28% of the domain names warned by LEMMINGS were re-registered at some point in time.

Also, domain names warned by LEMMINGS are re-registered faster than domain names not classified as being at risk. Figure 7 shows that the median time between release and re-registration for domains of the highest risk-category was 55 minutes, compared to 17 hours for domain names not warned.

Re-registering domain names does not automatically enable the new registrant to receive emails addressed to the previous registrant. First, the new registrant needs to add a MX record and point that MX record to a working mail server. We use our crawler to measure whether domain names of `.AnonCc` have a working mail server once per month. Here, we see that domain names warned by LEMMINGS have a higher chance of having a mail server assigned than other domain names. 40% of re-registered and warned domain names had a working mail server, compared to 35% of the other domain names.

The difference between warned domain names and others

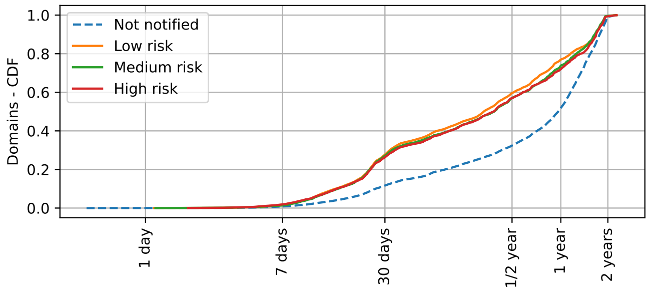

Fig. 8. Time between release out of quarantine and assignment of mail server by risk category.

	Alerted	Not alerted
Cancel-delete	355	719
No cancel-delete	54,055	524,805

TABLE IV
CONTINGENCY TABLE OF CANCEL-DELETE AND ALERTED.

Risk cat	Cancel Delete	%	Increase
Low	237	0.53%	3.7x
Medium	68	0.84%	6.1x
High	50	1.08%	7.9x

TABLE V
RESULTS PER RISK CATEGORY.

becomes larger if we look into how long it took until the new registrant assigned a working mail server to a domain name. Figure 8 shows that the median time between their release out of quarantine and the assignment of a working MX record is around 130 days for domain names warned by LEMMINGS compared to almost 1 year for other domains. Note that DMAP tests for working mail servers only once per month. For this reason, these numbers are a conservative estimation. In practice, new registrants might assign a mail server up to 30 days sooner.

C. Prevention of a Data Leak

It is not possible to directly measure the prevention of a data leak. In order to still be able to assess if LEMMINGS is able to prevent data leaks, we have selected three *proxy-metrics*. The first proxy-metric is based on the hypothesis that an increased usage of the domain registry cancel-delete request is based on a report by LEMMINGS. This could mean that the former registrant finds the warning useful and takes it seriously. The second hypothesis is that registrants, receiving a warning, could also inform potential senders (e.g. based on earlier communication) about the fact that they will delete a domain name. This can have an effect on the observed DNS queries – less mail attempts could lead to less MX queries. Our third metric relies on a survey, carried out among recipients of the LEMMINGS warning.

1) Cancel-deletes: The cancel-delete request is used to move a deleted domain name from the quarantine state back to the active state and is received and processed by our domain registration system.

To confirm our hypothesis that a warning by LEMMINGS leads to more cancel-delete request, we have created a baseline by analyzing the registry data for a 1-year period, from April 2021 until April 2022, before we started using LEMMINGS. For this period we have analyzed the number of domain name delete requests and the number of the cancel-delete request during the last 10 days of the quarantine period. This matches with the period in which LEMMINGS is active.

During the 12 month baseline period 627.285 domain name delete requests have been processed and 826 domain names received a cancel-delete request during the last 10 days of their quarantine period (0.13%).

Table IV shows the number of domain name registrations that received a cancel-delete request, and how many registrations had been alerted by our system. Based on this table, we can deduce that 0.647% of the alerted registrations received a cancel-delete request, while only 0.137% of the not-alerted registrations received such a request. A chi-square test of independence showed that this difference in cancel-delete ratio is significant, $\chi^2(1, N = 579,931) = 696.10, p < .001$.

The increase in cancel-delete requests for alerted domain name registrations is an indicator that the alert has helped the former domain name owners to take a informed decision to restore the domain. Another possible explanation could be that the domains that did not received an alert, are not that useful to their former owners and are therefore less likely to be restored from quarantine.

Table V shows that the cancel-delete ration increases along with the risk severity. This suggests that the probability a registrant will issue a cancel-delete request increases when the risk category of the alert increases.

Type of re-activated domain names By analyzing the web crawler data-set (Section IV-A), we are able to identify the web content category for a deleted, and no longer reachable, domain. Table VI shows the number of domain delete, alert and cancel-delete events, grouped by the content category. 21% of the cancel-delete domains, that have been restored to the active state from quarantine have been classified as business related domains. This while only 6.5% of all deleted domains have the same category. This could indicate that domain names that are of more value to the users have a higher chance of being recovered from quarantine. Table VII in the appendix contains a brief description for the other categories listed in Table VI.

Quarantine period We have a closer look at the distribution of cancel-delete requests received during the quarantine period (see Figure 9). The figure contains data points for a year long baseline period ending when LEMMINGS was introduced and for the period for which we evaluated LEMMINGS. The days of the 40-day long quarantine period are plotted on the x-axis, the percentage of cancel-delete requests received are on the

Category	**Cd**	**%Cd**	**Ad**	**%Ad**	**Dd**	**%Dd**
Business	75	21.13	8,917	16.39	40,249	6.85
Placeholder	68	19.15	7,813	14.36	126,383	21.50
Content	52	14.65	9,623	17.69	53,169	9.05
Parking	43	12.11	2,244	4.12	63,074	10.73
Ecommerce	23	6.48	3,922	7.21	20,002	3.40
Unknown	20	5.63	4,939	9.08	151,584	25.79
ServerDefault	19	5.35	5,127	9.42	34,088	5.80
ServerError	14	3.94	2,242	4.12	8,414	1.43
NotFound	11	3.10	3,485	6.41	32,096	5.46
Disallowed	10	2.82	2,324	4.27	26,206	4.46
Suspended	9	2.54	1,114	2.05	6,248	1.06
LowContent	7	1.97	1,656	3.04	17,357	2.95
OpenDirectory	2	0.56	204	0.37	759	0.13
NoContent	2	0.56	107	0.20	418	0.07
Captcha	0	0.00	50	0.09	151	0.03
Forum	0	0.00	9	0.02	47	0.01
NotReadyYet	0	0.00	381	0.70	2,353	0.40
ClientError	0	0.00	91	0.17	3,646	0.62
Unreachable	0	0.00	162	0.30	1,534	0.26

TABLE VI
RESULTS PER DOMAIN USAGE.

$^{\text{Category}}$ Classification of web content before domain was deleted.
$^{\text{Cd}}$ Number of cancel-delete request for domains.
$^{\text{Ad}}$ Number of alerted domains.
$^{\text{Dd}}$ Number of deleted domains.

y-axis. The vertical dashed line indicates the point in time (day=30) when LEMMINGS sends its alerts, We compared the 10-day period, after sending the alert, of the baseline quarantine period to the same period for LEMMINGS.

Most cancel-delete requests are received in the first thirteen days, after this the number of requests decreases gradually.

We see a request increase at the end of the LEMMINGS period, a similar increase is not found for the baseline. For the baseline, the total percentage of requests during the last 10-days was 6.2%, compared to 11.6% for LEMMINGS.

On day 28 we find an outlier in the LEMMINGS data, a spike of 2.9% (267 cancel-delete requests). After investigating, we concluded that 69% of the requests making up this spike are linked to the actions of a single registrar. The registrar executed 185 cancel-delete requests for domains that, based on the name, are all related to the same type of content and may have been deleted by accident.

2) Impact on DNS traffic: Taking a domain name out of quarantine is not the only option to prevent a data leak. In our mail to the registrants we also suggest them to notify their contacts about the deletion of the domain name. If these contacts stop trying to send emails to the deleted domain name then we would expect a decrease in the number of MX queries for that domain name.

In order to test this hypothesis, we take a subset of 20,965 domain names that entered quarantine between 2023-01-30 and 2023-02-06 and that were not taken out of quarantine before the quarantine period ended. For each domain name, we count the number of MX queries we receive starting 30 days before the quarantine period until 30 days after the domain name became available. We normalize the number of queries, such that a value of 1 stands for the highest of number queries received for a particular domain name per day.

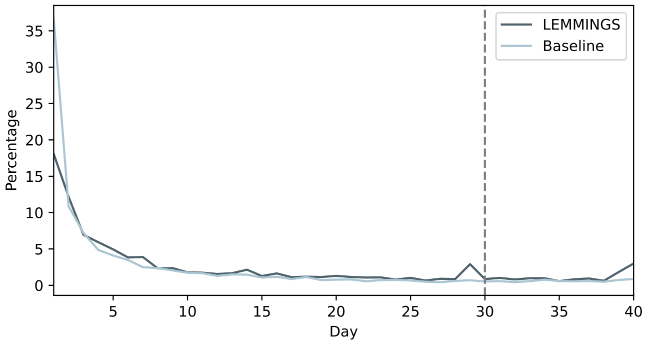

Fig. 9. Cancel-delete requests distribution during quarantine period.

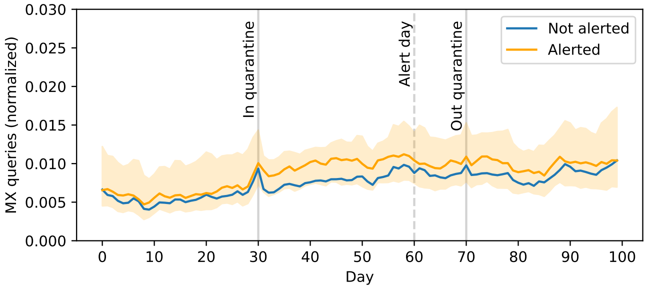

Fig. 10. Median number of MX queries for domain names alerted and domain names not alerted by LEMMINGS. The band marks the 25th and 75th percentile of queries for alerted domain names.

Figure 10 shows the normalized median of queries for domain names alerted by LEMMINGS and domain names that did not run a risk of a data leak. Here, we see an increase in queries when the domain names entered quarantine and on the day they became available. Also, we can observe a slight decrease in MX queries after 30 days of quarantine regardless of whether we alerted a domain name or not. This indicates that even if registrants took measures to reduce the number of emails we cannot see a measurable effect in our DNS traffic.

3) Survey: Since November 2022, we ask registrants to fill in an anonymous survey to help us to evaluate LEMMINGS. We include a link to the survey in the mail sent by LEMMINGS to the registrants. We do not collect any personal information during the survey. Until 2023-08-10, 166 registrants participated in this survey, while we warned registrants of 118,594 domain names in the same time period. Despite the low response rate, the survey helps us to get some understanding of the perception of LEMMINGS among registrants and its impact.

Figure 11 shows that the majority of participants perceive the warning by LEMMINGS as useful. 20% took action after receiving a warning and around the same share of participants think that LEMMINGS prevented a data leak (see Figure 12).

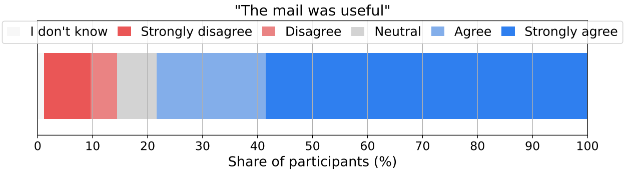

Fig. 11. Share of participants that agree or disagree with the statement "The mail was useful".

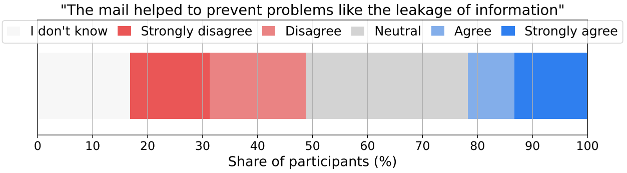

Fig. 12. Share of participants that agree or disagree with the statement "The mail helped to prevent problems like the leakage of information".

VI. Discussion and Future Work

Impact While we cannot point to a specific data leak that LEMMINGS prevented, our analysis shows that LEMMINGS warns those domain names that have a higher chance of having a data leak (Section V-B2), that the warnings reach the former registrants (Section V-C1), that the registrants find the warning useful and, according to some registrants, that it has prevented data leaks (Section V-C3).

When compared to the low number of complaints by registrants at our support desk and compared to the fact that only 6 registrars decided not to participate, we believe that LEMMINGS has largely achieved its goal.

Other measures against data leaks We would not need LEMMINGS if emails would always be encrypted, but the chance of encryption being used on a larger scale is low. The use of email encryption like PGP is perceived as too complicated for many and thus its adoption remains low [18].

A non-technical solution to the problem could be for registrants to hold on to domain names for a longer period of time. Depending on the number of domain names, the domain name themselves, and the TLD under which the domain names are registered, this might be costly. Also, one could argue that this would distort the domain name aftermarket.

Finally, since a few years, the Internet Corporation for Assigned Names and Numbers (ICANN) allows the registration of new TLDs. For some organizations, owning an own TLD might be feasible to tackle data leaks. Then, only authorized entities could then re-register a canceled domain. However, the costs and management-effort for such a private TLD might outweigh the risk of a potential data leak.

Contacting registrants We do not know how many registrants actually read our email. The low number of registrants retrieving their domain name from quarantine might be an indicator that email is not the most suitable channel to contact some registrants. This assumption has been confirmed by several studies that investigated effective vulnerability notifications [14], [21]. They found that email-based notifications might be perceived as not trustworthy or might not be delivered successfully at all.

In our case, we still believe that email is the most appropriate channel to reach out to registrants. This believe is based on, first, the fact that a panel of registrants found our email clear and trustworthy. Second, the fact that across a 30 day sample period only 5.9% of our emails could not be delivered by our email server. Third, the fact that the only other contact information registries have of their registrants are telephone number and postal address. We consider both too slow, impossible to automate on a larger scale, and even less reliable than email.

Unfiltered warnings Instead of trying to identify legitimate email attempts, we could also inform every registrant of a deleted domain name about the potential risk. This would have the advantage that we would not need elaborate filters and would guarantee that we would not miss any domain name running the risk of a data leak. However, as shown in §V, this imprecise approach could lead to large amounts of warnings sent to registrants and could even increase the chance that our warnings would be perceived as SPAM.

Future challenges and improvements In the future, we will develop LEMMINGS further, e.g. by improving our domain name filters and DNS filters. At the same time, we will also need to assess the impact of changes in the DNS on our approach.

Query name minimization (RFC 9156 [5]) is a DNS protocol extension that causes recursive resolvers to reduce the information shared with authoritative name severs to a bare minimum. This improves the privacy for end users and also means that operators of name servers of TLDs will only see queries for the NS (name server) and A and AAAA (IPv4 and IPv6 address) records. Magnusson et al. [17] showed that the adoption of RFC 9156 is already on the rise. At .AnonCc, we see that for the first week of September 2023 that 75% of larger resolvers[2] still send MX queries. However, this is already a reduction by 2% compared to the year before.

We will monitor the impact of query name minimization on LEMMINGS and explore other approaches (like machine learning) to identify DNS queries that are caused by mail transfers in the future.

VII. Ethics

Email may be used for sending confidential and sensitive information. For this reason, we have proceeded very carefully during the course of this study. Our first step was to create a privacy policy, in which we describe the aim of the project, the type of data used, the retention period and who has access the collected data. This privacy policy was then submitted to our internal privacy board which evaluated and approved the privacy policy.

[2] Resolvers that send at least 10,000 queries in this time period.

The study was limited to DNS traffic and website data collected by our crawler. The DNS traffic only contains metadata about the sender and recipient mail server infrastructure. We have no visibility into the actual mail transactions, for instance we do not know what email addresses are used and have no access to the mail contents.

When a deleted domain name exits the quarantine period, information identifying the registrant, such as the registrant name and email address, is deleted from the LEMMINGS database.

VIII. Related Work

Previous work discussed extensively the risk of phishing and other domain name related abuse. In this section, we focus on related work that specifically assesses the risk of expired domain names and work that discusses mail related data leaks.

Lauinger et al. [11] studied the expiration process at different TLDs and showed that many domain names are re-registered soon after deletion. These re-registered domain names can still have "residual trust", a concept introduced by Lever et al. [12] in 2016. They show that other services including email still try to use the expired domain names, which can threaten the security and privacy of users. They demonstrate, for example, that canceled domain names can lead to hostile takeovers of IP address space. The concept of residual trust has been followed up by other studies (e.g. [13], [4]), and in 2022 So et al. [20] took another deep dive and registered 201 domain names still receiving DNS queries. They showed that residual trust of domain names also affects Software-as-a-Service (SaaS), software libraries, and Internet radio and music stations – an observation that was confirmed by Liu et al. [16] in 2023.

Some studies took a closer a look at one of the threats described above. For example, Liu et al. [15] showed in 2016 that expired domain names of name servers are common and that they put domain names at risk of adversarial takeovers. Akiwate et al. [1] repeated this study on a larger scale in 2020 showing that the risk still remains. As one of the few, Hupkens et al. [8] focused solely on abuse of residual trust in email. They have registered 30 deleted domains and set-up email services on them. Then, they collected the incoming emails and found that there was sensitive data in emails belonging to six of these 30 emails.

Related to the threats above, but not misusing residual trust, Szurdi et al. [22] show that typo-squatting domains, impersonating domain names of popular services, can receive hundreds of thousands emails containing potentially sensitive information.

In contrast to our work, previous work mainly focused on the detection of domain names at risk. Our method has the goal to mitigate the risk at an early stage, that is, before an attacker could take over a domain name.

IX. Conclusions

In this paper, we propose an approach for preventing email-related data leaks caused by expired domain names. This approach can be applied at TLDs. We implemented this approach in our tool called LEMMINGS and deployed LEMMINGS at the ccTLD of *Anonymized EU Country*. Even though we cannot verify ourselves whether LEMMINGS has prevented data leaks, we showed that LEMMINGS is successful in warning registrants of potentially vulnerable domain names and that the registrants consider the warning useful. Thereby, we take a step towards preventing email-related data leaks on a large scale.

References

[1] Akiwate, G., Jonker, M., Sommese, R., Foster, I., Voelker, G.M., Savage, S., Claffy, K.: Unresolved Issues: Prevalence, Persistence, and Perils of Lame Delegations. In: Proceedings of the ACM Internet Measurement Conference. pp. 281–294 (2020)

[2] Anti-Phishing Working Group (APWG): APWG eCrime Exchange (eCX). https://apwg.org/ecx/ (2023), https://apwg.org/ecx/

[3] BNR Webredactie: Politiegeheimen liggen op straat (*In Dutch*). BNR (January 2017), https://www.bnr.nl/nieuws/binnenland/10316797/politiegeheimen-liggen-op-straat?disableUserNav=true

[4] Borgolte, K., Fiebig, T., Hao, S., Kruegel, C., Vigna, G.: Cloud Strife: Mitigating the Security Risks of Domain-Validated Certificates. In: Proceedings of the Applied Networking Research Workshop. p. 4. ANRW '18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3232755.3232859, https://doi.org/10.1145/3232755.3232859

[5] Bortzmeyer, S., Dolmans, R., Hoffman, P.: DNS Query Name Minimisation to Improve Privacy. RFC 9156, IETF (Nov 2021), http://tools.ietf.org/rfc/rfc9156.txt

[6] Daniël Verlaan: Jeugdzorg: Datalek dossiers kinderen Utrecht door 'foutje' in e-mail (*In Dutch*). RTL Nieuws (May 2019), https://www.rtlnieuws.nl/tech/artikel/4672826/jeugdzorg-datalek-dossiers-kinderen-utrecht-email

[7] Dzuba, E., Cash, J.: Introducing Cloudflare's 2023 phishing threats report. Tech. rep., Cloudflare (2023), https://blog.cloudflare.com/2023-phishing-report/

[8] Hupkens, J., Hodzelmans, S., Jansen, J., de Laat, C.: Analysis on mx-record queries of non-existent domains (2020)

[9] Lauinger, T., Buyukkayhan, A.S., Chaabane, A., Robertson, W., Kirda, E.: From Deletion to Re-Registration in Zero Seconds: Domain Registrar Behaviour During the Drop. In: Proceedings of the Internet Measurement Conference 2018. p. 322–328. IMC '18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3278532.3278560, https://doi.org/10.1145/3278532.3278560

[10] Lauinger, T., Chaabane, A., Buyukkayhan, A.S., Onarlioglu, K., Robertson, W.: Game of Registrars: An Empirical Analysis of Post-Expiration Domain Name Takeovers. In: USENIX Security Symposium. pp. 865–880 (2017)

[11] Lauinger, T., Onarlioglu, K., Chaabane, A., Robertson, W., Kirda, E.: WHOIS Lost in Translation: (Mis) Understanding Domain Name Expiration and Re-Registration. In: Proceedings of the 2016 Internet Measurement Conference. pp. 247–253 (2016)

[12] Lever, C., Walls, R., Nadji, Y., Dagon, D., McDaniel, P., Antonakakis, M.: Domain-Z: 28 Registrations Later Measuring the Exploitation of Residual Trust in Domains. In: 2016 IEEE Symposium on Security and Privacy (SP). pp. 691–706 (2016). https://doi.org/10.1109/SP.2016.47

[13] Lever, C., Walls, R.J., Nadji, Y., Dagon, D., McDaniel, P., Antonakakis, M.: Dawn of the Dead Domain: Measuring the Exploitation of Residual Trust in Domains. IEEE Security & Privacy **15**(2), 70–77 (2017)

[14] Li, F., Durumeric, Z., Czyz, J., Karami, M., Bailey, M., McCoy, D., Savage, S., Paxson, V.: You've got vulnerability: Exploring effective vulnerability notifications. In: 25th USENIX Security Symposium (USENIX Security 16). pp. 1033–1050 (2016)

[15] Liu, D., Hao, S., Wang, H.: All Your DNS Records Point to Us: Understanding the Security Threats of Dangling DNS Records. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. pp. 1414–1425 (2016)

[16] Liu, G., Jin, L., Hao, S., Zhang, Y., Liu, D., Stavrou, A., Wang, H.: Dial "N" for NXDomain: The Scale, Origin, and Security Implications of DNS Queries to Non-Existent Domains (2023)

[17] Magnusson, Jonathan and Müller, Moritz and Brunstrom, Anna and Pulls, Tobias: A second look at dns qname minimization. In: International Conference on Passive and Active Network Measurement. pp. 496–521. Springer (2023)

[18] Mauriés, J.R.P., Krol, K., Parkin, S., Abu-Salma, R., Sasse, M.A.: Dead on arrival: Recovering from fatal flaws in email encryption tools. In: The LASER Workshop: Learning from Authoritative Security Experiment Results (LASER 2017). pp. 49–57 (2017)

[19] Park, J., Khormali, A., Mohaisen, M., Mohaisen, A.: Where are you taking me? Behavioral analysis of open DNS resolvers. In: 2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN). pp. 493–504. IEEE (2019)

[20] So, J., Miramirkhani, N., Ferdman, M., Nikiforakis, N.: Domains Do Change Their Spots: Quantifying Potential Abuse of Residual Rrust. In: 2022 IEEE Symposium on Security and Privacy (SP). pp. 2130–2144. IEEE (2022)

[21] Stock, B., Pellegrino, G., Li, F., Backes, M., Rossow, C.: Didn?t You Hear Me? Towards More Successful Web Vulnerability Notifications. In: Proceedings of the 25th Annual Symposium on Network and Distributed System Security (NDSS '18). (February 2018), https://publications.cispa.saarland/1190/

[22] Szurdi, J., Christin, N.: Email typosquatting. In: Proceedings of the 2017 internet measurement conference. pp. 419–431 (2017)

[23] The Radicati Group Inc.: Email Statistics Report, 2022-2026. Tech. rep., The Radicati Group Inc (2022), https://www.radicati.com/wp/wp-content/uploads/2022/11/Email-Statistics-Report-2022-2026-Executive-Summary.pdf

[24] The Spamhaus Project - Frequently Asked Questions (FAQ): (2011), http://www.spamhaus.org/faq/answers.lasso?section=Spamhaus%20SBL#10

[25] Wullink, M., Moura, G.C., Müller, M., Hesselman, C.: ENTRADA: A high-performance network traffic data streaming warehouse. In: Network Operations and Management Symposium (NOMS), 2016 IEEE/IFIP. pp. 913–918. IEEE (Apr 2016)

A Crawler content category types

Category	Description
Business	Website used for business activities
Placeholder	Default website, created by service provider
Content	Generic content
Parking	Website is not used and is for sale
Ecommerce	E-commerce (shop) activity detected
Unknown	Unknown content
ServerDefault	Default ”Hello World” server page
ServerError	Server erros page
NotFound	Page is not found
Disallowed	Access to resource is denied
Suspended	The account linked to the website has been suspended
LowContent	Too little content found for classification
OpenDirectory	Website displays server directory listing
NoContent	No content found
Captcha	Only a captche is presented
Forum	A forum website has been found
NotReadyYet	Server is not yet ready to serve content
ClientError	Crawler experienced error during crawling
Unreachable	Website cannot be crawled

TABLE VII
Content type classifications by the web crawler.