From WHOIS to WHOWAS: A Large-Scale Measurement Study of Domain Registration Privacy under the GDPR

Chaoyi Lu*†, Baojun Liu*†¶✉, Yiming Zhang*†, Zhou Li§, Fenglu Zhang*, Haixin Duan*¶✉,
Ying Liu*, Joann Qiongna Chen§, Jinjin Liang$^{\|}$, Zaifeng Zhang$^{\|}$, Shuang Hao** and Min Yang††
*Tsinghua University, †Beijing National Research Center for Information Science and Technology,
{lcy17, zhangyim17, zfl20}@mails.tsinghua.edu.cn, {lbj, duanhx}@mail.tsinghua.edu.cn, liuying@cernet.edu.cn
§University of California, Irvine, {zhou.li, joann.chen}@uci.edu,
¶Qi An Xin Group, $^{\|}$360 Netlab, {liangjinjin, zhangzaifeng}@360.cn,
**University of Texas at Dallas, shao@utdallas.edu, ††Fudan University, m_yang@fudan.edu.cn

***Abstract*—When a domain is registered, information about the registrants and other related personnel is recorded by WHOIS databases owned by registrars or registries (called WHOIS providers jointly), which are open to public inquiries. However, due to the enforcement of the European Union's General Data Protection Regulation (GDPR), certain WHOIS data (i.e., the records about EEA, or the European Economic Area, registrants) needs to be redacted before being released to the public. Anecdotally, it was reported that actions have been taken by some WHOIS providers. Yet, so far there is no systematic study to quantify the changes made by the WHOIS providers in response to the GDPR, their strategies for data redaction and impact on other applications relying on WHOIS data.**

In this study, we report the first large-scale measurement study to answer these questions, in hopes of guiding the enforcement of the GDPR and identifying pitfalls during compliance. This study is made possible by analyzing a collection of 1.2 billion WHOIS records spanning two years. To automate the analysis tasks, we build a new system `GCChecker` based on unsupervised learning, which assigns a compliance score to a provider. Our findings of WHOIS GDPR compliance are multi-fold. To highlight a few, we discover that the GDPR has a profound impact on WHOIS, with over 85% surveyed large WHOIS providers redacting EEA records at scale. Surprisingly, over 60% large WHOIS data providers also redact non-EEA records. A variety of compliance flaws like incomplete redaction are also identified. The impact on security applications is prominent and redesign might be needed. We believe different communities (security, domain and legal) should work together to solve the issues for better WHOIS privacy and utility.

I. Introduction

The General Data Protection Regulation (GDPR) was established to set up new policies to protect the privacy of personal data within the European Union. Since it went into effect in May 2018, prominent changes have been made by companies across sectors to comply with the GDPR requirements. To measure the impact of the GDPR, previous works have focused on the web space, including website cookies [42], [39], online advertising [55], [96], [103], [102] and usability of privacy notices [104], [78], [79], [90], [50], [49], [27], [72].

Due to its broad scope, not only does the GDPR protect normal users browsing websites, users *setting up* websites and the associated infrastructure are also protected. One example is domain registration. After a user registers a domain name, e.g., `example.com`, its sponsoring registrar and upper-stream registry will store his/her personal information like name and address in the WHOIS database, and release it when receiving a WHOIS query. Such registration data is clearly within the GDPR's scope and in response, ICANN proposed a Temporary Specification [11] to instruct its contracted registries and registrars (or WHOIS providers)[1] to redact the personal information from WHOIS records.

However, redacting WHOIS records could also hamper the utility of applications that protect Internet users. For a long time, WHOIS serves as a critical data source for the security community, providing clues to track malicious domain owners and the associated cyber-attack activities. Disagreements between legal authorities and technical communities [73] on how to align WHOIS data with the new privacy regulations were raised. Anecdotally, some investigators complained about the utility loss of WHOIS data, and that the time to trace cybercrime has been significantly elongated [25]. By contrast, some people insist that tracing threats using the post-GDPR WHOIS information is still good enough [26]. Despite those anecdotal evidence, so far there is not yet a systemic study *quantifying* the impact of the GDPR on the WHOIS system, answering questions like *how many WHOIS providers redact WHOIS data? how do they redact the data? how large is the impact on the security applications?* These knowledge gaps should be filled so as to guide the enforcement of privacy policies (including the GDPR and others like the CCPA [13]) in the future, which motivates us to carry out this study.

Challenges. Analyzing WHOIS data in the lens of the GDPR is non-trivial, and several challenges need to be addressed ahead. 1) The domain ecosystem is very fragmented: there are thousands of registries and registrars running WHOIS, resulting in inconsistent data format and wide-spread data sources. 2) The time when WHOIS providers complied with the GDPR is never announced. Thus, WHOIS records covering

Network and Distributed Systems Security (NDSS) Symposium 2021
21-24 February 2021, San Diego, CA, USA
ISBN 1-891562-66-5
https://dx.doi.org/10.14722/ndss.2021.23134
www.ndss-symposium.org

[1] We use WHOIS providers to refer to both registrars and registries.

a long time-span (e.g., months before and after the GDPR effective date) need to be collected for in-depth analysis of the responses of WHOIS providers. 3) Due to the vagueness of the ICANN Temporary Specification, WHOIS providers can apply various redaction methods. Therefore, applying simple methods like keyword matching on WHOIS records to check their compliance will result in high error rates.

Our Study. In this paper, we report the first comprehensive data-driven analysis on the GDPR compliance of the domain registration ecosystem. To address the first and second challenges, we collaborate with an industrial partner and access a passive WHOIS dataset containing WHOIS records collected from **Jan 2018 to Dec 2019**. We are able to analyze **1.2 billion WHOIS records** about **267 million domain names** in total. Not only are the changes before and after the GDPR effective date observed, a large number of EEA (the European Economic Area) domains are covered (over 32 million). To address the third challenge, We design a system named `GCChecker` based on unsupervised learning and natural language processing (NLP). Our key insight is that a GDPR-compliant WHOIS provider prefers to use simple and automated approaches to replace records at scale. Therefore, for each WHOIS provider, by analyzing the statistical distribution of its record values, we can conclude whether it complies with the GDPR and the level of compliance. We use DBSCAN to identify outliers (non-compliant records) and a NER annotator to refine the results. We find the outlier ratio can indicate the degree of GDPR-compliance at high accuracy.

Major findings. We run `GCChecker` on the entire passive WHOIS dataset and highlight the major findings below. 1) The enforcement of the GDPR has brought a significant impact on the WHOIS ecosystem: **over 85% large WHOIS providers** (in terms of sponsored EEA domains) we study (89 registrars and 54 registries) are now GDPR-compliant, meaning that the WHOIS fields containing personal information are redacted at scale. Surprisingly, 3 registries are still not fully-compliant as of Dec 2019, though they are direct delegates of ICANN. To understand the impact on smaller WHOIS providers, we adjust the parameters of `GCChecker` and include additional 48 registrars and 65 registries. As a result, we find smaller providers are more likely to be partially-compliant or non-compliant. Besides, we discover various flawed implementations of GDPR compliance. For example, 6 registrars mask only part of the registrant's fields, and 21 registrars do not offer alternative channels to contact domain holders, which are actually requested by ICANN. Regarding the scope of protected domains, we find in a surprise that **over 60%** large WHOIS providers apply the same protection mechanism on both EEA and non-EEA domains, though only EEA domains are regulated by the GDPR.

Our measurement results indicate fundamental changes of the WHOIS system in response to the GDPR. Given that alternative channels for security researchers and practitioners (e.g., tiered-access system) are not well-maintained (e.g., requests to view WHOIS data are usually rejected [25]), we expect many security applications have to be re-designed. To quantify such impact, we survey 51 security papers published at five conferences in the past 15 years that leverage WHOIS data. Among them, **69%** surveyed papers need to use redacted WHOIS information. We believe both the security and domain community should work closely together to address the accessibility issue of WHOIS data as soon as possible.

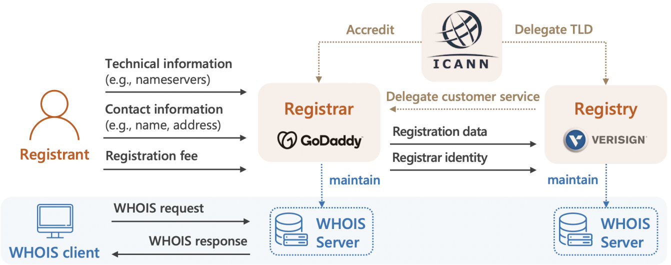


Fig. 1: The domain registration hierarchy and WHOIS

Finally, we have been reporting our findings to providers containing non-compliant WHOIS records. We also discussed with ICANN staff and a few registrars about the causes behind their reactions to the GDPR. The lessons we learned are mainly two: 1) the vagueness of ICANN's instructions and the short preparation time window (the ICANN Temporary Specification was released only 1 week before the GDPR effective date) forced many WHOIS providers to take the "*safest*" approach and sanitize all records blindly; 2) the lack of checking tools caused many flawed implementations. The lessons suggest enforcing privacy policies is still a complex task, requiring more efficient collaboration across communities. To contribute to the communities, we develop an online checking tool based on `GCChecker` and plan to release it in the near future.

Contributions. The contributions are listed as follows:

- *New methodology.* We design a new system named `GCChecker`, for automated GDPR-compliance check.
- *Measurement findings.* We analyze 256 WHOIS providers and assess their compliance status. Implementation issues that need to be addressed are identified.
- *Online checking tool.* We develop an online checking tool (at `https://whoisgdprcompliance.info`) for WHOIS providers to check their compliance status.

II. Background

In this section, we provide background of the domain registration hierarchy and WHOIS database. We then introduce basic legal requirements of the GDPR, as well as its impact on the current WHOIS system.

A. Domain Registration (WHOIS) Database

As shown in Figure 1, the domain name space is managed by registries and registrars in a hierarchical structure. The Internet Corporation for Assigned Names and Numbers (ICANN) creates Top-Level Domains (TLDs, e.g., `.com`) and delegates them to *registries* (e.g., Verisign) which operate the TLD zones. Registries then delegate customer service to *registrars* (e.g., GoDaddy) which sell domain names to *registrants* (or domain holders). All registrars and registries of generic TLDs (gTLD) are contracted with ICANN under the Registrar Accreditation Agreement (RAA) [2] and Registry Agreement (RA) [6].

Domain registration data. As required by policies of the RAA, registrars collect and retain in their databases both *technical information* (e.g., name of the authoritative servers)

TABLE I: Registration data publishing requirements of the ICANN Temporary Specification [11]

Registration Data Fields	Data Subjects	Data Publishing Requirements
Name, Street, City, Postal Code, Phone, Fax	Registrant, Admin, Tech, Other	Unless provided consent from the registrant, a) provide a **redacted value** (substantially similar to "redacted for privacy"), or b) prior to RDAP[1] implementation, provide **no information** of, or **not publish** the field.
Organization, State/Province, Country	Admin, Tech, Other	
Email address	Registrant, Admin, Tech	For registrars only: provide an **anonymized email address** or **web form**, which should **facilitate email communication** with the data subject.

[1] RDAP [77] is designed as the successor protocol of WHOIS. But in the short term, WHOIS will not be replaced [9].

and *contact information* (e.g., registrant information) of their sponsoring domains. Specifically, contact information includes the *registrant*'s name, postal address, email address and telephone number, as well as the *administrative contact* (the agent appointed by the registrant or his/her company), *technical contact* (the person responsible for maintaining the authoritative servers) and *billing contact* (the person responsible for paying the domain's renewal fees). When a Second-Level Domain (SLD) is registered, the registrar also submits a copy of the registration data to upper-level registries in a model called "thick WHOIS" [7], unless the SLD is under three TLDs: `.com`, `.net` and `.jobs`. For the three TLDs, contact information is only retained by registrars (the model is called "thin WHOIS"), but they are expected to move to "thick WHOIS" by the end of 2020 [12]. According to their contracts with ICANN, both registrars and registries should offer *free query-based access* to their registration databases.

WHOIS: the lookup protocol of registration data. RFC 3912 [40] specifies the WHOIS protocol as the standard interface to query the domain registration database. To look up the registration data (or WHOIS record) of a domain, a WHOIS client sends a TCP request with the domain name to port 43 of WHOIS servers of the domain's sponsoring registrar and registry. Alternatively, the client user can also visit the web interfaces of WHOIS providers to fetch the WHOIS record. WHOIS data is maintained in a semi-structured textual format but the format is inconsistent across WHOIS providers [70], which makes it challenging to parse at scale. Section III-A describes how we handle the collected WHOIS records.

B. General Data Protection Regulation (GDPR)

The General Data Protection Regulation (EU) 2016/679 [5], or GDPR, is a data protection regulation designed to "harmonize" privacy laws of the EU member countries. Recital 6 [4] says the GDPR aims to provide a *high-level framework about protecting personal data* when the data flows within the (European) Union and out to other countries. Repealing the former Directive 95/46/EC [1], the GDPR was adopted in April 2016 and officially went into effect on 25 May 2018. Below we highlight its key legal requirements.

Processing personal data. Article 4 of the GDPR defines *personal data* as "any information relating to an identified or identifiable natural person". As a result, names, location data and online identifiers (e.g., email addresses, IP addresses and browser cookies) are considered as personal data. It also defines *processing* as "any operation or set of operations which is performed on personal data", including collection, storage and disclosure.

Consent from user. Article 6 of the GDPR ensures the rights of data subjects in controlling their data. In particular, data subjects can "give consent to the processing of his or her personal data for one or more specific purposes". Note that data protection is enforced by default, thus silence from the data subjects means no consent [8].

Territorial scope. Article 3 of the GDPR defines its territorial scope, which can be expanded outside the EU. Specifically, the GDPR applies to "the processing of personal data of data subjects who are in the (European) Union", *regardless of whether the processing takes place in the (European) Union, or whether the processor locates in the (European) Union.*

C. GDPR's Impact on Domain Registration

Due to its wide protection scope, the GDPR has introduced a profound impact on Internet applications relying on personal data. For example, websites are required to ask for explicit consent before setting browser cookies [42]. Because domain registration collects information of registrants and other personnel, it is within the scope of the GDPR.

ICANN Temporary Specification. To fill the gap between the GDPR's high-level requirement and low-level implementation of data protection, ICANN released a Temporary Specification for gTLD Registration Data [11] on 17 May 2018, which is one week before the GDPR effective date. The document applies to *all gTLD registry operators and ICANN-accredited registrars*, and aims to maintain the accessibility of the current WHOIS system "*to the greatest extent possible*". It retains the current registration data collection procedure, thus domain holders still provide their personal information to registrars. However, a WHOIS provider must take additional steps when releasing domain registration data (e.g., replying to WHOIS queries), if it is: 1) located in the European Economic Area (EEA), 2) located outside the EEA but offers registration services to registrants in the EEA, or 3) engaging a data processor in the EEA. Table I provides a summary of the requirements, covering data subjects of registrants, admin, tech and other[2]. Particularly, before the GDPR, WHOIS privacy protection services (e.g., WhoisGuard [15]) have been used by plenty of registrars to hide registrants' information from spammers, marketing firms and online fraudsters. Such services install an anonymous proxy identity for a registrant in the WHOIS database. When all fields in Table I are masked by the proxy, no additional changes are needed to comply with GDPR.

Regarding the scope, the ICANN Temporary Specification gives WHOIS providers flexibility to choose whether the

[2] Billing contact is not mentioned in the specification (it is different from "other"). We find it is rarely released in WHOIS records, so we neglect it in the following analysis.

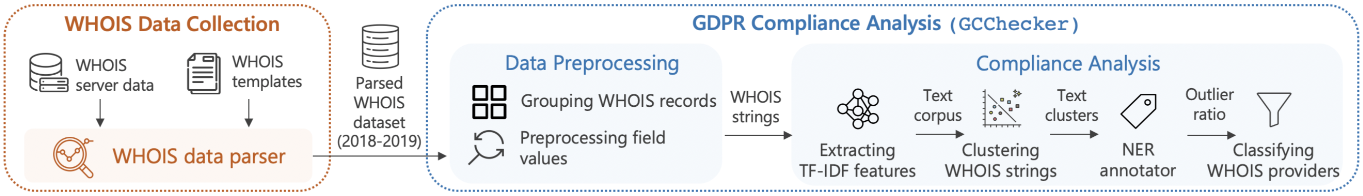


Fig. 2: Overview of WHOIS data collection and GCChecker

protection applies in a global basis or GDPR-governed regions only. In other words, it is acceptable if a provider chooses to release the original WHOIS data of domain holders living outside the EEA. In the remainder of this paper, we use EEA domains and EEA records to refer to domains registered by EEA registrants and WHOIS records associated with these domains. Non-EEA domains and records are defined similarly.

III. Methodology

In this study, we aim to assess at scale whether WHOIS providers comply with the GDPR, in particular the ICANN Temporary Specification. We are also interested in how data protection is actually enforced. However, answering those research questions is non-trivial, as WHOIS data is not well structured and WHOIS providers can apply any data redaction method. In this section, we elaborate on our methodology of WHOIS data collection and GDPR compliance analysis (i.e., `GCChecker`). Figure 2 overviews the key steps.

A. *WHOIS Data Collection*

We aim to provide a longitudinal (before and after the GDPR effective date) and latitudinal (covering a wide range of WHOIS providers) view of the GDPR's impact on domain registration. To this end, we collaborate with an Internet security company and leverage its historical WHOIS dataset for this study. The company maintains a passive DNS service (similar to Farsight DNSDB [48]) for threat hunting, which aggregates DNS requests and responses logged by affiliated DNS resolvers across regions. When a domain name is newly observed (i.e., being queried by an Internet user), the system will attempt to fetch its WHOIS record. For domain names under TLDs using the "thin WHOIS" model (i.e., `.com`, `.net` and `.jobs`), their WHOIS records are collected from registrars. For domains under other TLDs, the WHOIS records are collected from registries. Occasionally, for example when domains are about to expire, the system re-fetches their WHOIS records to obtain updates. The WHOIS data collection system has been in operation since 2016, and we use the data spanning **from Jan 2018 to Dec 2019 (2 years)**.

Parsing WHOIS records. The standard document of WHOIS [40] only specifies its transport mechanisms, yet in practice providers do not agree on the format of WHOIS records [34]. The lack of consensus significantly hampers large-scale analysis of WHOIS data [70], especially for TLDs adopting the "thin WHOIS" model. To address this issue, open-source and commercial WHOIS parsers have proposed template-based (e.g., Ruby Whois [10]), rule-based (e.g., pythonwhois [17]) and statistical approaches (e.g., [70]). Our industrial partner uses a template-based method, where hundreds of WHOIS templates are manually created for different WHOIS providers. The templates are also regularly reviewed to accommodate format changes. Figure 3(a) shows a snippet of a parsed WHOIS record.

```
// (a) Parsed WHOIS record in JSON
{
    “collected_time”: 2018-01-01T04:00:00Z, “domain_name”: example.com
    “domain_creation_date”: 1995-08-14T04:00:00Z,
    ”whois_server”: whois.iana.org, “iana_id”: 376,

    “registrant_name”: Example Name, “registrant_street”: 123 Example Rd
    “registrant_city”: Example Town, “registrant_postalcode”: 00000
    “registrant_phone”: +1.00000000, “registrant_country”: United States,
    “registrant_email”: 1a79a4d60de6718e8e5b326e338ae533@example.com
}
```

```
// (b) WHOIS String
(registrar_376, non-EEA, registrant, week 2018-01-01)    // main key
example name \t 123 example rd \t example town \t 00000 \t +1.00000000 \t
32 \t example.com \t none \t 1
```

Fig. 3: (a) A WHOIS record after parsing (only technical and registrant fields are shown), (b) the WHOIS string after preprocessing (Steps I and II described in Section III-B).

Based on our definitions of (non-)EEA records (in Section II), a WHOIS record is identified as a (non-)EEA record, if its registrant country is one of the (non-)EEA countries. The ICANN Temporary Specification does not suggest redaction of registrant country information, and we find that most WHOIS providers are following this rule based on empirical analysis[3]. Before the next steps, we remove all WHOIS records without registrant country information (e.g., being redacted or having format errors, which cover around 12.7% records).

Table II presents the statistics of our parsed WHOIS dataset, after removing records without registrant country information. The dataset contains **1.2 billion WHOIS records of 267 million domain names**. Both newly-registered and older domains (e.g., 13% domains are created before 2010) are included. Around 12% domains are EEA domains and over 67% records are collected from registrar WHOIS servers.

Limitations. Our industrial partner collects WHOIS records of domains observed in its passive DNS dataset, which might be biased due to the geo-location of the affiliated resolvers. While we acknowledge this limitation, our evaluation shows that the dataset has a satisfactory coverage of domain names globally (219 countries and 12% EEA domains) and a wide range of TLDs (783 in total). We believe the results obtained from this dataset are representative. Another limitation is that the records are only collected from WHOIS servers by querying port 43. In the meantime, web interfaces of providers are not examined and there is a chance that WHOIS data is not sanitized there. However, web-based WHOIS are often protected by CAPTCHA which prevents data collection [54], [64], thus in this paper we focus on port-43 WHOIS services.

[3] We check all 194.6M WHOIS records collected in Jan, Apr, Jul and Oct 2019, which belong to 726 WHOIS providers. Only 10 providers (e.g., the `.name` registry) prevent us to extract registrant country from any of their WHOIS record. Among 632 (87%) providers we extract registrant country information from over 90% records.

TABLE II: Statistics of the parsed WHOIS dataset

Year	Count of		Domain Creation Date			Domain TLD				Registrant Region			Data Source	
	Record	Domain	~'09	'10 ~'17	'18 ~'19	#	.com	.net	.org	#	EEA	non-EEA	Registrars	Registries
2018	659,184,231	211,614,203	15.7%	58.5%	25.8%	758	64.2%	6.58%	5.28%	218	12.9%	87.1%	64.7%	35.3%
2019	583,179,357	215,772,034	14.5%	42.4%	44.1%	754	66.5%	6.19%	4.71%	219	12.4%	87.5%	70.1%	29.9%
Total	**1,242,363,588**	**267,634,833**	**13.4%**	**49.7%**	**36.9%**	**783**	**63.8%**	**6.19%**	**4.74%**	**219**	**12.2%**	**87.8%**	**67.2%**	**32.8%**

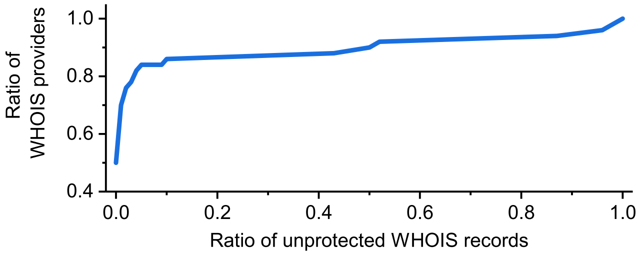


Fig. 4: Cumulative distribution of 50 selected WHOIS providers and ratio of unprotected WHOIS records in D_G.

B. GDPR Compliance Analysis

The major goal of this study is to assess whether the *EEA records* released by WHOIS providers follow the requirements of the GDPR and ICANN Temporary Specification. To this end, we take the EEA records collected from each provider as a whole and analyze its degree of compliance.

Dataset for empirical analysis. We inspect a sample of 50 WHOIS providers (40 registrars and 10 registries) to gain insights into how they process WHOIS data. The providers are selected by their share of registered domains (see Appendix A, they account for over 50% of the total share) and by having a large number of EEA records (over 1,000 records collected per month). For each provider, we randomly sample 1,000 EEA records collected in Dec 2019 (50,000 records in total) and manually label each record by whether its contact information (i.e., *all* fields listed in Table I) is protected. This dataset (termed as D_G) serves as ground truth for our system design.

Among the 50,000 records in D_G, 5,647 (11.3%) are labeled as unprotected. Figure 4 shows the cumulative distribution of the 50 providers' ratio of unprotected records in D_G. We find a knee around 5% unprotected records – 84% (42 of 50) providers have a lower ratio, and we can spot clear data protection measures in their released records. Later we use this observation to assign three compliance levels to WHOIS providers (described in Step IV).

Technical Challenge. To protect personal data, instructions of the ICANN Temporary Specification are not strict: it asks the redacted fields to be replaced by values *substantially similar* to "redacted for privacy", instead of requiring the use of the exact string. From D_G we also find redacted fields under different wording and languages, which deters automated textual analysis. Below are some examples:

- "privacidad WHOIS" (in Spanish)
- "obfuscated WHOIS"
- "statutory masking enabled"

As a result, learning whether a WHOIS field is redacted under the GDPR becomes a difficult task. Simply matching keywords in each field will lead to a high error rate.

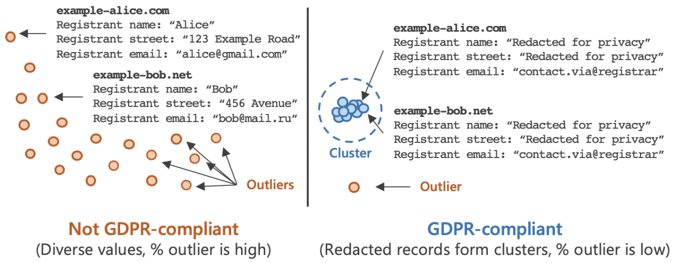

example-alice.com
Registrant name: "Alice"
Registrant street: "123 Example Road"
Registrant email: "alice@gmail.com"
example-bob.net
Registrant name: "Bob"
Registrant street: "456 Avenue"
Registrant email: "bob@mail.ru"
Outliers
Not GDPR-compliant
(Diverse values, % outlier is high)
example-alice.com
Registrant name: "Redacted for privacy"
Registrant street: "Redacted for privacy"
Registrant email: "contact.via@registrar"
example-bob.net
Registrant name: "Redacted for privacy"
Registrant street: "Redacted for privacy"
Registrant email: "contact.via@registrar"
Cluster
Outlier
GDPR-compliant
(Redacted records form clusters, % outlier is low)

Fig. 5: Outlier ratio and GDPR compliance. The ratio is an indicator of WHOIS records containing unprotected data.

Insight on WHOIS textual similarity. While scanning each individual WHOIS record and checking its compliance is error-prone and time-consuming, we find that by computing the statistical distribution of WHOIS record values per provider, we can tell whether it complies with the GDPR and the level of compliance. Our key observation from D_G is that a GDPR-compliant WHOIS provider prefers to use simple and automated approaches to replace record values at scale, resulting in high homogeneity of the values, especially on EEA domains. Otherwise, the values are expected to be diversified because they contain information of different registrants. As a result, the redacted records should form big *clusters* with zero or a small number of *outliers* (i.e., samples that do not belong to any cluster). Figure 5 illustrates this insight, where each dot represents a WHOIS record after vectorization. Inspired by this observation, we leverage *clustering algorithm* and use *outlier ratio* as an indicator of GDPR compliance. We find that this approach works generally well but sometimes a domain holder can register a bulk of domains which can also form a cluster. We apply a lightweight NLP-based approach to filter such clusters (described in Step III). We design `GCChecker` based on this insight, which contains the following steps.

Step I: Grouping WHOIS records. We first group the WHOIS records according to their providers. If the provider is a registry, we group them by their *WHOIS servers*[4]. If the provider is a registrar, we group them by their *registrar IDs*. For each accredited registrar, ICANN assigns a unique numerical ID [18], which is included in every WHOIS record (see Figure 3, field "`iana_id`"). In this step we also remove records under drop-catch registrars (e.g., ID-1756[5] DropCatch.com LLC and ID-1008 SnapNames, LLC). The main reason is that drop-catch registrars will put their domains on auction or proactively register domains without getting a back-order [61]. As a result, their WHOIS records can be intentionally filled with similar values (e.g., "domain on sale") without any relation to a real registrant.

[4] We obtain a list of registry WHOIS servers from the IANA Root Zone Database [19]. The field "`whois_server`" of parsed WHOIS records is checked against the list.

[5] We use ID-1756 to refer to the registrar whose ID is 1756.

Next, we separate each group by time windows (e.g., weeks) to study its changing dynamics. The WHOIS records within each window are further grouped according to their registrants' regions (i.e., EEA and non-EEA) and data subjects (i.e., registrant, admin and tech contacts). In the end, each WHOIS record is associated with a *main key*, and records under different keys are analyzed separately. The main key is a quadruple:

$$(provider, registrant\_region, data\_subject, time\_window)$$

Under a main key, we only fill in the fields that are relevant to the embedded data subject. Note that in each time window we only use the "current" version of WHOIS records (i.e., records collected in this time window) to identify their registrant region and provider. Because WHOIS records in each time window are analyzed separately, the system is not affected by the drift of domain ownership (e.g., domain transferring).

Step II: Preprocessing field values. Before clustering WHOIS records and assessing their homogeneity, we preprocess the fields of each record to make clustering more efficient and effective. In detail, we concatenate the values of fields that should be masked (e.g., registrant name, phone and email, listed in Table I) using tab characters to produce a single string and convert it to lower case. Figure 3(b) shows an example of the output string, which we call WHOIS string. If the value of a field is empty, we fill it with a dummy value ("`none`").

We also need to take extra care of *pseudonymized values* which might be GDPR-compliant. In contrast to anonymized strings (e.g., "redacted for privacy"), pseudonymized values are uniquely generated for different data subjects. For instance, a provider could use `[hash_1]@example.com` and `[hash_2]@example.com` (where `[hash_1]` and `[hash_2]` are different hash strings) to mask two email addresses. However, WHOIS strings with pseudonymized values are hardly clustered together, which increases the outlier ratio and affects the accuracy. To eliminate its impact, based on manual analysis on D_G, we apply the following rules to handle pseudonymized values.

- *Domain name in redacted values.* If a redacted value contains the domain name itself (e.g., "owner of `example.com`"), we replace the domain name with a fixed string "`domain`".
- *Number in registrant name.* If the registrant name contains a digital number (e.g., "customer no. 123456"), we replace the number with a fixed string "`number`".
- *Email address.* As summarized in Table I, the email address field can be replaced by a pseudonymized email address or a hyperlink to a web form. Given that the pseudonymized values under a WHOIS provider tend to be generated automatically (e.g., using hash values of the same length), we use a quadruple template to represent this field, including 1) the length of the local-part of the email address; 2) the domain of the email address; 3) the domain of the link to the web form; 4) the number of phrases in this field, separated by white space (some registrars fill multiple phrases in this field). In Figure 3's example (`1a79a4d60de6718e8e5b326e338ae533@example.com`), the length of the local-part is 32, and the domain name of the email address is `example.com`. The value is not a web link, thus the third part is filled with "`none`". Finally, the number of phrases is 1.

The set of WHOIS strings with the same main key can be considered as a text corpus, and we use TF-IDF [94] to compute their term frequencies as features. TF-IDF is widely used to generate statistical features for document clustering and we use it for a similar purpose. We use white spaces, tab characters and punctuation as separators to split terms.

Step III: Clustering WHOIS strings. The WHOIS strings generated in Step II comprise all fields that should be protected, thus we choose to cluster them as a whole for efficiency. We cluster WHOIS strings based on the TF-IDF features and compute the *outlier ratio* to infer the degree of GDPR compliance for a WHOIS provider. We leverage DBSCAN [46], a density-based clustering algorithm which can detect outliers, for the task. DBSCAN treats clusters as high-density areas separated by low-density areas. It does not require the number of clusters to be specified ahead, which can find arbitrarily-shaped clusters and scale to large datasets. For WHOIS strings under the same main key, we use DBSCAN to mark the outliers and calculate their ratio over the number of total records. The only parameter of DBSCAN is `min_samples` which specifies the minimum size of a cluster, and we empirically set it to 25.

For GDPR-compliant providers, we expect that protected WHOIS records are similar and thus clustered together. However, domains with the same contact information (e.g., registered in bulk by the same registrant) can also form clusters, which lowers the outlier ratio. Such clusters are different from "GDPR-compliant clusters" in that personal information is included. Therefore, we leverage the NER (Named-entity Recognition) annotator of the Stanford CoreNLP Natural Language Processing Toolkit [71] to find clusters containing information of natural persons. The toolkit is based on trained CRF (Conditional Random Field) sequence taggers and a system for processing temporal expressions, and can recognize a given named entity as PERSON, LOCATION, ORGANIZATION and MISC. In our setting, if one sample of WHOIS strings within a cluster contains names labeled as "PERSON", we label all records in the cluster as outliers. Though the time consumed by applying CoreNLP on a record is non-negligible, the overall overhead is small thanks to the clustering process executed beforehand.

To enable large-scale analysis, we implement the clustering module with MapReduce [41] and scikit-learn [82], and run the program on a Hadoop cluster. The WHOIS strings are preprocessed by mappers, allocated to reducers according to their main keys, and clustered by all reducers in parallel. Due to the memory limit (2GB) of each reducer machine, we only keep a random sample of 20,000 records under each main key. On average, the clustering task of each main key finishes in 3 minutes, and it takes 25 minutes to analyze one-week data of all WHOIS providers.

Step IV: Classifying WHOIS providers. With the outlier score computed for each main key value (different provider, region, data subject and week), we classify a WHOIS provider by the level of GDPR compliance. The outlier scores under each provider will be compared against a set of thresholds and we leverage the ground-truth dataset (i.e., D_G) to determine

TABLE III: A sample of WHOIS providers and outlier ratios (results of registrant fields of EEA records).

Compliance Degree	WHOIS Provider	Weekly Outlier Ratio				
		2018			2019	
		Jan 01	Apr 02	Aug 06	Mar 04	Oct 07
Fully	ID-146 GoDaddy.com, LLC	0.901	0.894	0.002	0.001	0.000
	ID-69 Tucows Domains Inc.	0.942	0.955	0.012	0.002	0.001
	ID-2 Network Solutions, LLC	0.953	0.966	0.001	0.000	0.001
Partially	ID-81 Gandi SAS	0.642	0.652	0.117	0.114	0.107
Not	ID-1068 NameCheap, Inc.	0.721	0.461	0.548	0.777	0.868
	.us Registry	1.000	0.879	0.473	0.871	0.886

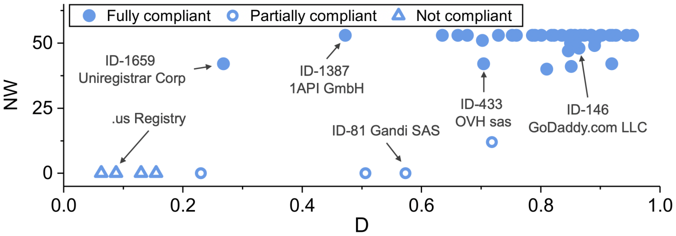


Fig. 6: Distribution of WHOIS providers under NW and D (results of registrant fields of EEA records).

their values. From our prior observations in Figure 4, we assign three compliance levels to the WHOIS providers:

- *Fully-compliant.* Over 95% WHOIS records are redacted[6]. 42 providers fall into this category.
- *Partially-compliant.* 50%-95% WHOIS records are redacted. 4 providers fall into this category.
- *Not compliant.* Less than 50% WHOIS records are redacted. 4 providers fall into this category.

Next, we generate the weekly outlier scores of the 50 providers on the entire 104 week's data (from Jan 2018 to Dec 2019). Table III shows a sample of providers and weeks. The degree of changes differs significantly by the three categories before and after the GDPR enforcement deadline, indicating the effectiveness of clustering on the larger dataset. However, using one outlier score to classify a provider is not enough, due to the sampling done by our industrial provider and `GCChecker`. We experiment with different statistical metrics on the sequence of outlier ratios and find two that are most distinguishing: 1) the *number of weeks* (NW) in 2019 when the outlier ratio is below 0.05, and 2) the *drop* (D) of the average outlier ratios before May 2018 and after May 2018. In Figure 6, we plot NW and D of the 50 providers and find the ranges are largely different by the three categories. In the end, we set two conditions to classify a provider by comparing NW and D to thresholds:

- **Condition 1:** If NW is over 40, the WHOIS provider is categorized as "fully-compliant".
- **Condition 2:** When Condition 1 is not satisfied, if D is over 0.2, the WHOIS provider is categorized as "partially-compliant", otherwise categorized as "not compliant".

C. System Evaluation

In this section we first evaluate the key components of `GCChecker` separately, then its end-to-end effectiveness.

Data preprocessing rules. In Step II, three data preprocessing rules are generated from D_G to handle pseudonymized field values, and here we assess their generality across other WHOIS providers. To this end, we use the entire dataset collected in Dec 2019 and randomly sample 20 EEA records under each provider for manual inspection, resulting in 10,200 WHOIS records from 510 providers (including the 50 providers in D_G). We find no additional pseudonymized values that should be preprocessed.

Clustering and NER annotator. In Step III, DBSCAN and CoreNLP are used to mark unprotected WHOIS records as outliers. To evaluate the performance of this module, we run the program on D_G where 5,647 records have been manually labeled as unprotected. The system reports 4,691 outlying records, in which 4,620 are also manually labeled as unprotected, so precision is 98.4% (4,620/4,691) and recall is 81.8% (4,620/5,647). The false positives are resulted from unpopular choices of redacted values, while false negatives are large domain holders not correctly recognized by CoreNLP.

While the high precision ensures most GDPR-compliant records can be identified, the recall is less satisfactory. A potential effect of false negatives is that providers may receive lower outlier ratios than the actual ratio of unprotected records, and we assess the distribution of errors on results of D_G. For the 42 fully-compliant providers, we find they only have clusters of protected WHOIS records (which should not be outliers), so they are not affected by false negatives. By contrast, 6 in 8 partially- and non-compliant providers have clusters of large domain holders that are not identified as outliers, and their outlier ratio can be lowered by 0.07 to 0.38 (0.17 on average). However, considering the compliance degrees we set, 49 of 50 providers (including 5 providers which receive a lowered outlier ratio) still get an outlier ratio in the range of their corresponding categories. Only one non-compliant provider receives an outlier ratio of 0.49 (lowered by 0.38 because of false negatives), which falls into the range of partially-compliant. This provider can be correctly classified as not compliant, if we compare its sequence of weekly outlier ratio against the conditions in Step IV, as in other weeks the provider is given outlier ratios in the correct range. Therefore, the impact of false negatives on the final result is expected to be insignificant.

End-to-end effectiveness. To evaluate the end-to-end effectiveness of `GCChecker` (i.e., whether the output WHOIS provider category is correct), we compile a test set (termed as D_T) of 20 WHOIS providers, including 10 registrars and 10 registries. The 20 providers are randomly selected beyond the 50 providers in D_G (but they should also have enough EEA records for analysis)[7]. Similarly, for each provider we sample 1,000 EEA records collected in Dec 2019, manually label each record, and give its level of GDPR compliance based on the ratio of unprotected records. We label 17 providers as fully-compliant (e.g., ID-1239), 1 as partially-compliant (ID-1725)

[6] According to the ICANN Temporary Specification, domain holders may consent WHOIS providers to release their real contact information. From the behavior of 84% providers in D_G, we consider 5% unprotected records as a conservative upper bound of "fully-compliant".

[7] Because in D_G we select the top 50 providers, the 20 providers in D_T are smaller in domain share.

and 2 as not compliant (e.g., ID-52). Further, the clustering results on D_T show that only the one partially-compliant provider receives a lowered outlier ratio by 0.08 because of false negatives, but the output ratio (0.24) is still in the correct range (i.e., 0.05 to 0.50 for partially-compliant providers). We then run `GCChecker` on the 2-year dataset to generate the weekly outlier ratio for each provider and compare them against our conditions. All 20 providers are correctly classified by `GCChecker`, suggesting that the system performs well end-to-end.

Choice of system parameters. There are several parameters in the design of `GCChecker` and here we discuss our choices on them. In Step II, DBSCAN takes `min_samples` as the minimum size of a cluster. A high value results in more outliers, which potentially removes more unpopular choices of redacted values from clusters. Also, fewer providers can be analyzed because more EEA records are required for each week. By contrast, a low value results in more clusters, including those of large domain holders, which affects the accuracy. We set this value to 25 and our evaluation shows that the system generally works well end-to-end, while ensuring that 143 providers can be analyzed in a weekly basis (in Section IV).

In Step IV, the conditions classifying WHOIS providers are established based on thresholds (i.e., NW and D). The thresholds are selected from clustering results of providers in D_G on the 104-week dataset (see Figure 6). Combined with the provider categories we set, a WHOIS provider is classified as fully-compliant only if the outlier ratio is stable enough at a low level (i.e., remain lower than 0.05 for a long period, through NW).

Limitations. 1) We are unable to directly run `GCChecker` on WHOIS providers with a small number of EEA records in the weekly time window (e.g., ID-420 Alibaba Cloud based in China). For long-tail providers, we loosen the time window to 2 months to aggregate more EEA domains for clustering, and provide a separate analysis in Section V. By this change, we are able to significantly increase the number of providers to be assessed. Though finer-grained dynamics are missing for this analysis (e.g., how providers act right before and after the GDPR enforcement deadline), the general degree of compliance can be learned. 2) The recall of outlier detection (81.8%) could be improved. Identifying whether a WHOIS record is protected is challenging, due to a lack of context in the WHOIS records. We use the NER-based method which is considered as the best-effort approach by the NLP community. Combined with DBSCAN, we find that the system generally works well end-to-end based on evaluations. 3) The GDPR and ICANN Temporary Specification give EEA registrants rights to allow WHOIS providers to publish their real contacts. We cannot identify whether the unprotected records are consented solely from data analysis. As we choose 5% unprotected records as a conservative upper bound for "fully-compliant", providers with a large number of EEA users allowing to publish their real data could be classified as otherwise. For providers not classified as fully-compliant by `GCChecker`, we have been reporting the results to them and list some of the feedback in Section IV. We also release an online tool for WHOIS providers to check their compliance status (discussed in Section VII).

D. Ethical Considerations

The major ethical considerations of this study are the collection of WHOIS records and the analysis of (non-redacted) personal data inside the WHOIS dataset. To avoid overloading the WHOIS servers, our industrial partner enforces strict rate limit when sending WHOIS queries. We are informed by our industrial partner that they have not received warnings from any WHOIS provider so far. The same data collection method has also been used by previous works [70], [62], [61], [63] which collect and parse WHOIS data at scale. For instance, WHOIS records of 102M domains were crawled by [70] and the full WHOIS dataset from DomainTools is used by a study in 2020 [63]. Regarding data analysis, the WHOIS dataset is provided to us for research purposes only, and we execute all programs on the servers of the industrial partner. During our study, the WHOIS records in the dataset are never shared with third parties. We also consulted professional lawyers about legal issues of analyzing EEA records and are notified that it does not violate the GDPR regulations. As our data processing is for research purpose only, with no relation to offering goods or services to EU citizens, analyzing WHOIS data is allowed according to Recital 23 [3] of the GDPR.

IV. WHOIS Providers with Large Number of EEA Domains

In this section, we study the WHOIS providers maintaining a large number of EEA domains. We run `GCChecker` on the weekly data to carry out macro-level (e.g., how they comply with the GDPR in general) and micro-level analysis (e.g., how they mask the WHOIS records).

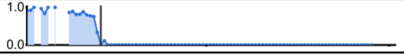
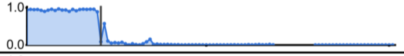
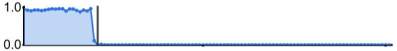
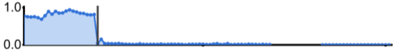
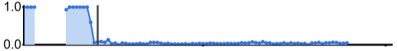
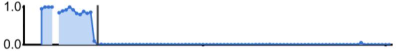
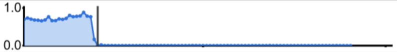
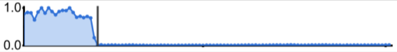
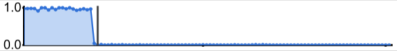
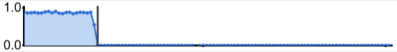
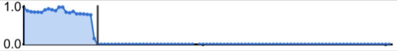
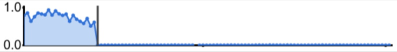
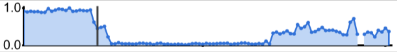
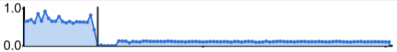
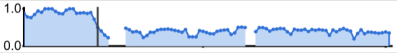
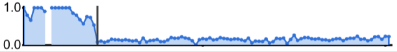
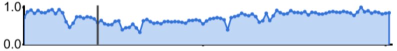
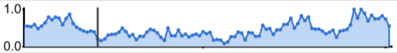
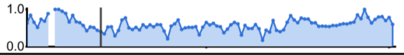
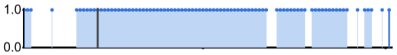
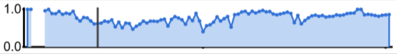
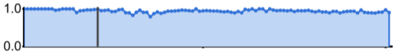
A. WHOIS Provider Selection

We use the number of EEA records observed in our dataset per week to select qualified providers for this measurement task. In particular, we count the number of weeks where over 50 EEA records (i.e., 2 × `min_samples`) are collected for a provider, and choose the providers whose week numbers are over 90 (i.e., 90% of all weeks). This selection method ensures our clustering method can be executed without change and sufficient weeks can be measured. In total, **89 registrars** and **54 registries** meet the criteria, and Table IV shows a subset. According to ICANN's reports, the selected registrars sponsor **63.08%** of all registered domain names (see Appendix A for details of registrar domain share). For registries, as ICANN do not report their share, we show the number of sponsoring TLDs in our dataset. Note that leading registries like VeriSign (managing `.com` and `.net`) are not included because their managed TLDs use "thin WHOIS" and are thus not queried by our industrial partner. As the selected WHOIS providers offer services to EEA registrants, we consider them to be under the scope of the GDPR and ICANN Temporary Specification.

B. Status of GDPR Compliance

For each WHOIS provider, we apply `GCChecker` on the WHOIS strings under the *registrant contact of EEA domains* to obtain their weekly outlier ratios and their compliance categories (i.e., "fully", "partially" or "not"). We neglect tech and admin contact as we find the trends are similar.

TABLE IV: GDPR compliance analysis results of WHOIS providers with large number of EEA domains (2018-2019)

Category	WHOIS Provider	Name	CC	Share / # TLD [1]	WHOIS Server	Trending of Weekly Outlier Ratio (Registrant fields of EEA records) [2]	Contact Masking [3]			Email Anonymization	
							Redact	Empty	PriPro	Interface	Forward [4]
Fully compliant count: 124	**Registrar** count: 73 share: 54.23%	ID-146 GoDaddy.com, LLC	US	29.1%	whois.godaddy.com		○	●	○	Web	●
		ID-69 Tucows Domains Inc.	CA	4.68%	whois.tucows.com		●	○	○	Web	○
		ID-2 Network Solutions, LLC	US	3.31%	whois.networksolutions.com		●	○	○	Mail	○
		ID-48 eNom, LLC	CA	2.64%	whois.enom.com		●	○	○	Web	●
		ID-895 Google LLC	US	1.94%	whois.google.com		○	●	○	Mail	●
		ID-440 Wild West Domains, LLC	US	1.29%	whois.wildwestdomains.com		○	●	○	Web	●
		ID-433 OVH sas	FR	1.05%	whois.ovh.com		○	●	○	Mail	●
		ID-625 Name.com, Inc.	US	0.93%	whois.name.com		●	○	○	Web	●
		ID-9 Register.com, Inc.	US	0.82%	whois.register.com		●	○	○	Mail	○
	Registry count: 51	Public Interest Registry (PIR)	US	1 (.org)	whois.pir.org		○	●	○	--	--
		Donuts Inc.	US	230	whois.donuts.co		●	○	○	--	--
		XYZ.COM LLC	US	1 (.xyz)	whois.nic.xyz		○	●	○	--	--
Partially compliant count: 9	**Registrar** count: 8 share: 1.76%	ID-269 Key-Systems GmbH	DE	0.66%	whois.rrpproxy.net		○	●	●	Mail	●
		ID-81 Gandi SAS	FR	0.65%	whois.gandi.net		●	○	○	Mail	●
		ID-2487 Internet Domain Service BS Corp	BS	0.16%	whois.internet.bs		●	○	○	Mail	●
	Registry count: 1	Afilias, Inc.	US	28	whois.afilias-srs.net		○	●	○	--	--
Not compliant count: 10	**Registrar** count: 8 share: 7.09%	ID-1068 NameCheap, Inc.	US	4.46%	whois.namecheap.com		○	○	●	Mail	●
		ID-1479 NameSilo, LLC	US	1.59%	whois.namesilo.com		○	○	●	Mail	●
		ID-472 Dynadot, LLC	US	0.93%	whois.dynadot.com		○	○	●	Mail	●
		ID-52 Deluxe Small Business Sales, Inc. d/b/a Aplus.net	US	0.06%	whois.names4ever.com		○	○	○	○	○
	Registry count: 2	NeuStar, Inc.	US	1 (.us)	whois.nic.us		○	○	○	--	--
		Fundacio puntCAT	ES	1 (.cat)	whois.cat		○	○	○	--	--

(●: found / supported; ○: not found / not supported; --: not applicable)

[1] For registrars, show domain share; for registries, show the number of sponsoring TLDs observed in our dataset.

[2] The vertical line in the graph shows the GDPR effective date. We remove weeks where the number of WHOIS records do not reach 50 (the lower limit).

[3] Redact / Empty / PriPro: using redacted values (e.g., "redacted for privacy") / empty values / privacy protection services to mask contact information.

[4] Forward: whether the email anonymization interface supports forwarding messages to domain registrants. Applicable to registrars only.

Observation 1: The enforcement of the GDPR has profound impact on WHOIS data release: over 85% large WHOIS providers we studied are now GDPR-compliant.

Table IV presents the detailed clustering results of the top WHOIS providers. In total, 124 (86.7% of 143) providers are classified as fully-compliant, including 73 registrars and 51 registries. In Sections IV-C and IV-D we will further investigate their time of GDPR compliance and protection measures. Meanwhile, 9 (6.3% of 143) providers are classified as partially-compliant and the remaining 10 (7.0% of 143) providers are classified as not compliant.

Observation 2: Not all registries are fully GDPR-compliant as of Dec 2019.

Though registries are supposed to be fully-compliant as it works closely with ICANN, surprisingly, we find 3 exceptions. Two registries (NeuStar, Inc. and Fundacio PuntCAT) of two TLDs (`.us` and `.cat`) are classified as not compliant and over 90% of their WHOIS records are outliers. `.us` is a particularly interesting case. According to a decision by the US National Telecommunications and Information Administration (NTIA) in 2005, the information of `.us` domain holders should not be kept private [74]. On the other hand, we believe the decision

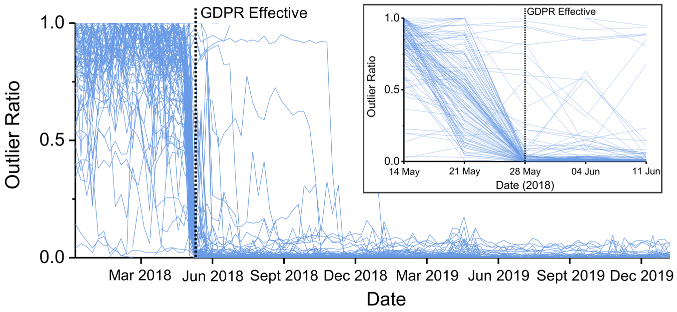


Fig. 7: Weekly outlier ratio of 124 GDPR-compliant providers (results of registrant fields of EEA records).

of NTIA should be revisited since `.us` does offer registration services to EEA residents, who are protected by the GDPR. One registry (Afilias, Inc.) is classified as partially-compliant. We find that the domain outliers are all under the `.srl` TLD, while other WHOIS records are protected.

> **Observation 3**: While most registrars are striving to protect their domains, flawed implementations are discovered.

By manually inspecting the outlying records, we find some providers do not mask all contact fields required by the ICANN Temporary Specification. For instance, we find that 4 registrars (ID-2487, ID-447, ID-1011 and ID-1564) do not protect the address fields (i.e., Registrant Street, City and Postal Code). For around 10% domains sponsored by ID-81 and ID-1725, only the email address field is masked. We recommend the providers to update their data protection policies.

While we pinpoint the root causes of compliance failure of 6 providers, we are not able to identify obvious reasons for the remaining ones solely from WHOIS data. We have been reporting the issue to the providers. In the feedback from NameCheap and NameSilo, their explanation is that registrants opt-in to display their WHOIS information (e.g., through email verification). In the feedback from Gandi, the unprotected information comes from registrants who chose to opt-out the default privacy protection service before the GDPR went effective, and their choice is respected. To help WHOIS providers address this issue, we also implement an online tool to check their current status of GDPR compliance, leveraging some building blocks of `GCChecker` (described in Section VII).

C. Timeline of GDPR Compliance

> **Observation 4**: Over 80% large GDPR-compliant WHOIS providers complete data protection timely before the GDPR went effective (on 25 May 2018), but they waited to take actions after the adoption of the ICANN Temporary Specification (on 17 May 2018).

Figure 7 shows the weekly outlier ratio of 124 GDPR-compliant WHOIS providers altogether. Around the GDPR effective date (25 May 2018) we observe a significant drop in the outlier ratio of most providers, suggesting that large-scale data masking is performed. To quantify this change, for each provider we use the first week where the outlier ratio drops below 0.05 as the starting time of compliance. As a result, 100 (80.6% of 124) WHOIS providers completed their data protection in time before the GDPR went effective. 11 providers (e.g., ID-1659 Uniregistrar Corp) show a delayed compliance of more than one month, i.e., later than Jun 2018.

When taking a closer look at the timeline around GDPR compliance, we find that prominent actions were taken *after* the adoption of the ICANN Temporary Specification on 17 May 2018 for most providers. In other words, data masking is performed *within only one week*. Before 17 May 2018, we only find 2 registrars (ID-1001 Domeneshop AS dba domainnameshop.com and ID-1666 OpenTLD B.V.) taking measures at scale. This finding resonates with a previous study on web privacy showing that over 70% privacy policy changes happened close to the GDPR effective date [42].

Though there is a 2-year grace period for organizations to prepare for the GDPR, it still takes nearly 2 years for registrars and registries to take necessary actions at scale. We discussed our observation with a large international registrar and learned that WHOIS providers prefer to wait until the ICANN Temporary Specification was released due to a void of specific guidance. Moreover, as the specification only left one week for the providers to make changes before the GDPR effective date, they tend to choose simplistic data masking strategies, resulting in changes beyond EEA records (further discussed in Observation 6).

D. Data Protection Measures

As shown in Table I, the ICANN Temporary Specification includes guidance on contact masking (for fields other than email address) and email anonymization. To understand how WHOIS providers follow the guidance, we inspect the clusters of WHOIS records of the GDPR-compliant providers and characterize their types of measures.

Contact masking. The ICANN Temporary Specification suggests that contact information can be either redacted or filled with an empty value. In practice, this suggestion is widely adopted by WHOIS providers: 46 registrars (e.g., ID-69 Tucows Domains Inc.) and 12 registries (e.g. `.vip` and `.amsterdam` registries) use redacted values; 24 registrars (e.g., ID-146 GoDaddy.com, LLC) and 39 registries (e.g., `.org` and `.site` registries) use empty values. Below are examples of redacted values used by WHOIS providers:

- “redacted for privacy” (e.g., ID-69 Tucows Domains Inc.)
- “statutory masking enabled” (e.g., ID-2 Network Solutions, LLC)
- “non-public data” (e.g., ID-625 Name.com, Inc.)
- “not disclosed” (e.g., ID-1505 Gransy, s.r.o.)
- “redacted” (e.g., `.wien` Registry)

Besides, WHOIS privacy protection services are also leveraged by registrars for contact masking. In this case, real registrant information is replaced with the name and address of the information of the service. While typically the services are paid, we find that they are made free by some registrars in response to the GDPR (e.g., WhoisGuard [15]). By identifying a small set of keywords (e.g., “privacy” and “protected”), we find 13 registrars mask a portion of records, and 3 registrars (e.g., ID-1456 NetArt Registrar Sp. z o.o.) mask all records using WHOIS privacy services to comply with the GDPR.

TABLE V: Examples of email anonymization interfaces

Interface	ID	Registrar	Value of Masked Email Addresses
Web (40 IDs)	146	GoDaddy.com, LLC	https://www.godaddy.com/whois/results.aspx?domain=***.com
	440	Wild West Domains, LLC	https://www.secureserver.net/whois?plid=1387&domain=***.com
	625	Name.com, Inc.	https://www.name.com/contact-domain-whois/***.com/registrant
	1659	Uniregistrar Corp	https://uniregistry.com/whois/contact/***.com?landerid=whois
	151	PSI-USA, Inc.	https://contact.domain-robot.org/***.com
Email (12 IDs)	895	Google LLC	f***************7@proxyregistrant.email (valid for 5 days)
	433	OVH SAS	g******************j@n.o-w-o.info
	291	DNC Holdings, Inc.	***.com-registrant@directnicwhoiscompliance.com
	1443	Vautron Rechenzentrum AG	a********q@domprivacy.de
	74	Online SAS	3**************9.1*****9@spamfree.bookmyname.com
Others (15 IDs)	69	Tucows Domains Inc.	https://tieredaccess.com/contact/0******d-8**0-4**5-9**2-d***3
	2	Network Solutions, LLC	abuse@web.com
	48	eNom, LLC	https://tieredaccess.com/contact/0******d-8**0-4**5-9**2-d***3
	9	Register.com, Inc.	abuse@web.com
	141	Cronon AG	domaincontact@reg.xlink.net

> **Observation 5**: Though most GDPR-compliant registrars offer direct communication channels to domain holders after email anonymization, over 25% do not offer such channel.

Email anonymization. Though email addresses are instructed to be anonymized, redacting them or making them empty is not recommended. Domain holders need to be reached via email for various reasons, such as domain validation of TLS certificates [31], vulnerability notification [66], [97], [89] and inquiries of domain reselling. As required by the ICANN Temporary Specification, registrars should set up interfaces that *facilitate direct communication* with the domain holder (see Table I, the requirement applies to registrars only). Among the 73 GDPR-compliant registrars, we find that over 70% registrars are following the requirements: 40 leverage web links and 12 use pseudonymized email addresses which are unique for each domain. Another 15 registrars avoid direct messaging – instead, they use tiered access systems (e.g., `https://tieredaccess.com` used by 6 registrars) or unified email addresses (e.g., `abuse@web.com`) as proxies to hide registrants' emails. Some examples are shown in Table V. The remaining 6 registrars (e.g., ID-140 Acens Technologies, S.L.U.) redact email addresses together with other contact information, which is also not recommended.

To learn how the interfaces are operated, we perform field study on top registrars (15 using web links and 5 using pseudonymized email addresses). We register domain names as holders under the registrars and send messages via the web links or pseudonymized email addresses. It turns out that the interfaces simply forward our messages to the registrant's real email address. Therefore, the sender's email address can be found in the `From` or `Reply-to` header fields of the received message, and the domain holder must use his/her *real email address* to reply to the message. We consider the protection offered by the interfaces insufficient and suggest the registrars 1) set up a mail transfer agent (MTA) [110] for automatic email forwarding and 2) configure the mail server to sanitize headers for better privacy protection.

> **Observation 6**: Though the GDPR is supposed to regulate EEA data only, over 60% providers also sanitize non-EEA WHOIS records, causing a global impact to WHOIS.

Scope of protection. As described in Section II, a WHOIS provider may choose to apply data protection to EEA domains only or beyond. To learn the preferences of WHOIS providers, we select a subset of the 124 GDPR-compliant providers, of which each one also has over 50 non-EEA records collected per week for at least 90 weeks, to measure the difference between EEA and non-EEA records. There are 88 providers matching the criteria and we plot the weekly outlier ratios of EEA and non-EEA records in Figure 8 of the top 16. Note that we plot separate lines for registrant, admin and tech, and find the general trends of them are similar. Surprisingly, we find that 80 providers (of 124, 64.5%) choose to aggressively sanitize non-EEA domains as well. The remaining 8 providers are all registrars (e.g., ID-2 Network Solutions, LLC and ID-625 Name.com, Inc.), offering protection to EEA domains only. As a result, though EEA domains constitute a small share (only 12% in our dataset, as shown in Table II), they do impact the entire WHOIS system due to the GDPR.

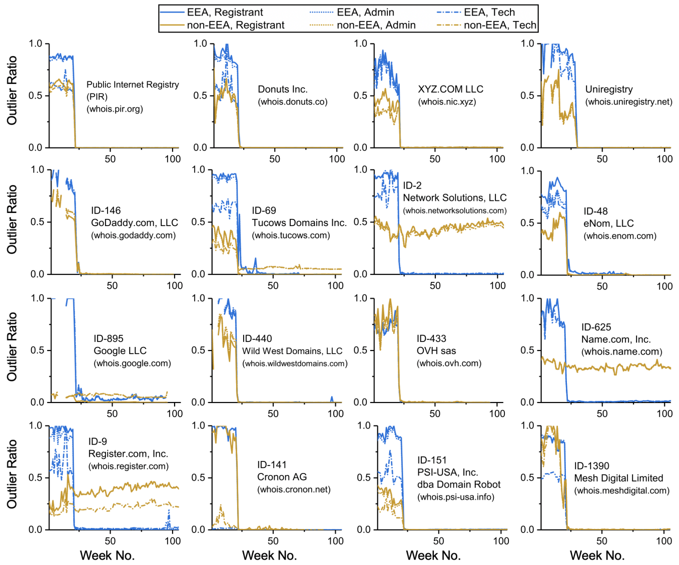
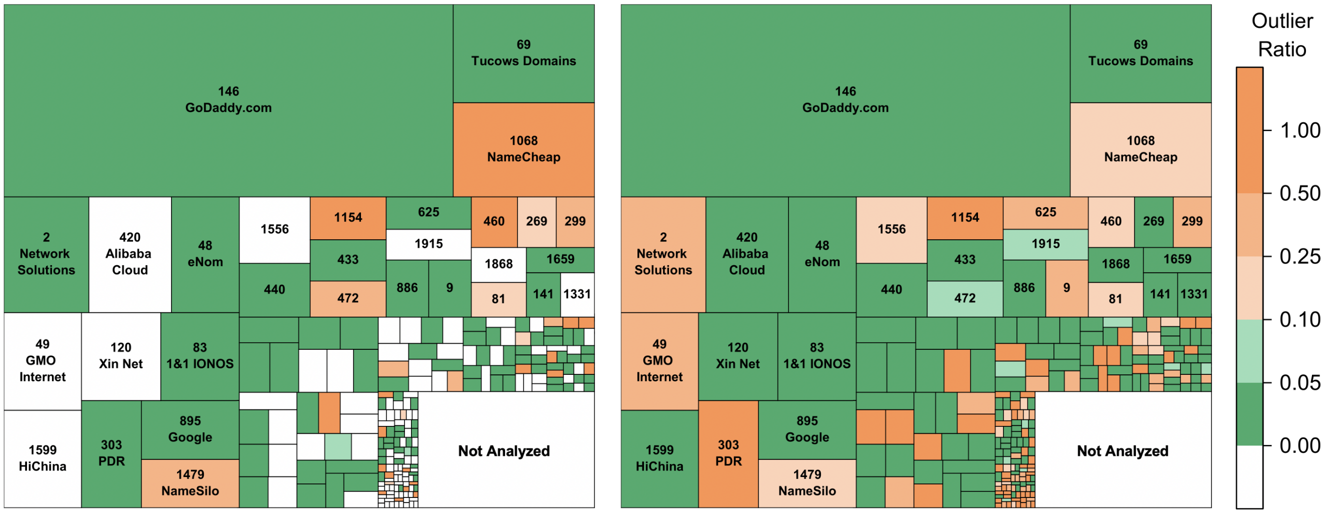


Fig. 8: Comparison of weekly outlier ratio, separated by data subjects (registrant, admin and tech) and registrant region (EEA and non-EEA).

To understand the rationale behind applying protection to non-EEA records, we inquired a large registrar and were told that dealing with EEA and non-EEA records together is easier than treating them separately. Because there was only one week for the providers to respond after the ICANN Temporary Specification, they find it challenging to identify which data is governed by the GDPR, and are thus forced to use the "*safest*" solution to protect all records. In addition, they are concerned about other new regional privacy laws (e.g., the California Consumer Privacy Act [13]), so changing all records at once saves extra work in the future. On the other hand, changing WHOIS records universally could have an adverse impact on security applications relying on WHOIS data, and we discuss this issue in Section VI.

V. Long-Tail WHOIS Providers

To provide a broader view of the status of GDPR compliance, we extend the time window from 1 week to *2 months* in order to cover WHOIS providers with smaller number of EEA domains. A provider is selected for this measurement task if it has over 50 (i.e., 2 × `min_samples`) EEA records collected in *every 2-month window* (12 windows in total). As a result, we are able to inspect **256 WHOIS providers**, increasing the number under the previous task (143) by nearly 80%. Accordingly, Condition 1 is adjusted: if the outlier ratio stays



(a) EEA domains

(b) non-EEA domains

Fig. 9: The outlier ratio of registrars in the last 2-month window (Nov - Dec 2019). The size of each block indicates the registrar domain share. Registrars which have less than 50 records collected in this period are not analyzed.

below 0.05 for 5 windows in 2019, the provider is classified as fully-compliant. Below we describe the measurement results on registries and registrars separately.

Registries. In total, **119 registries** are selected and we find that 113 (95.0% of 119) are classified as fully-compliant. Except for the 3 registries already discussed in Section IV (`.us`, `.cat` and `.srl`), we find 3 more country-code TLD (ccTLD) registries using flawed data protection measures. In detail, registries of `.gs` and `.cx` are not protecting the registrant address fields for around 50% of their sponsoring domains. Meanwhile, the `.mn` WHOIS server (`whois.nic.mn`) copies records from lower-level registrar WHOIS servers and the flaw is rooted in the registrars. Though ccTLD registries are not contracted with ICANN under the RA, which are thus not required to enforce the ICANN Temporary Specification, we recommend them to review their current policies when domains can be registered by EEA citizens.

Observation 7: Registrars with larger domain share have better compliance with the GDPR. The status is more worrying for smaller registrars.

Registrars. In total, **137 registrars** are analyzed, with a total domain share of 72.77%. Similar to our prior observations in Section IV, we find that most providers have responded to the GDPR well: 105 (76.6% of 137, with a total domain share of 60.76%) are classified as fully-compliant. Meanwhile, we also find that registrars with larger domain share tend to comply with the requirements better: 21 out of 32 partially- or non-compliant registrars have a domain share of less than 0.07%. In Figure 9 we visualize the relation between the domain share and compliance level for EEA and non-EEA domains separately. The size of each block indicates the domain share of a registrar, and the color marks the outlier ratio. We choose the last 2-month time window (Nov - Dec 2019) to focus on the most recent status. One interesting observation is that for large registrars sponsoring a small number of EEA domains, data masking is also extensively applied to non-EEA domains (e.g., ID-420 Alibaba Cloud).

VI. WHOIS Usage in Security Applications

WHOIS data has been a key ingredient in powering many security applications, such as domain reputation systems and spam detection systems. However, due to the data redaction performed by WHOIS providers, the effectiveness of these applications becomes questionable. This issue has been discussed [59], [56] but no study has quantified the impact. In this section, we make the first attempt from the perspective of security literature.

A. *Survey of Security Literature*

The high-level idea of this task is to collect academic papers describing the usage of WHOIS and classify them based on how WHOIS is used. We focus on academic papers because most of them have a clear description of the features and the papers are easier to collect. For systems developed by the industry, based on our discussion with a number of security companies, WHOIS also provides key features (e.g., [67]) for applications like threat intelligence.

Methodology for paper collection. As the first step, we download all research papers published at 4 top-tier security conferences (NDSS, USENIX Security, IEEE S&P, ACM CCS) and 1 leading measurement conference (ACM IMC) since 2005, and select those using WHOIS data. To fetch papers, we build a web crawler based on Chromium [14], collect conference program pages and extract paper downloading links. On the corpus of 4,304 downloaded papers, we manually build a list of keywords (e.g., “WHOIS” and “domain”) and search in the papers to filter out irrelevant works. For the 193 remaining papers, we read all of them and remove false positives (e.g., papers using the IP WHOIS database).

Observation 8: The GDPR’s impact on security applications relying on WHOIS could be profound, as 69% of surveyed papers need to use redacted information.

In the end, we are able to find **51 papers** using WHOIS data. Among them, **35 (69%)** use fields that should be redacted

TABLE VI: Security literature that use redacted WHOIS data

Category	Reference	WHOIS Field			Usage	Details
		R	A, T	Em		
Domain Security	Paxson13 [81]	◐	◐		V	Identify DNS tunnel usage
	Alrwais14 [23]	◐	◐	◐	D, M	Identify infiltration targets Categorize domains
	Halvorson15 [51]	◐		◐	M	Infer domain ownership
	Vissers15 [108]	●	●	●	M	Infer domain ownership
	Plohmann16 [83]	●		●	M	Analyze usage of DGA domains
	Chen16 [36]			●	M	Identify domains registrants
	Vissers17 [107]	●		●	D	Attack vectors
	Liu17 [69]	◐	◐		M	Infer domain ownership
	Alowaisheq19 [21]	◐		◐	M	Validate malicious reuse
	Sivakorn19 [95]	●			D	Features for detection
	Le Pochat20 [63]	●		●	D	Features for detection
Spam Scam Fraud	Christin10 [37]	●	●	●	M	Group miscreants
	Reaves16 [86]	●			M	Identify phishing campaigns
	Miramirkhani17 [75]			●	M	Group scam domains
	Kharraz18 [58]	●	●	●	M	Group domain owners
	Bashir19 [28]	◐		●	M	Group publishers
Cybercrime	Wang13 [109]	◐		◐	M	Infer domain ownership
	Khan15 [57]	◐		◐	D	Cluster adversarial typosquatting
	Du16 [44]	●		●	M	Infer domain ownership
	Yang17 [112]			●	M	Analyze underground organizations
Privacy	Zimmeck17 [113]	◐	◐		M	Identify cross-device trackers
	Ren18 [87]	●		●	D	Identify vulnerable domains
	Vallina19 [105]	◐	◐		M	Infer website owners
HTTPS Certificates	Delignat-Lavaud14 [43]	◐	◐		V	Infer domain ownership
	Cangialosi16 [33]			●	M	Infer domain ownership
	Roberts19 [88]			●	D	Infer domain ownership
Mobile Security	Alrawi19 [22]	◐	◐	◐	D, N	Label data Report vulnerabilities
	Van Ede20 [106]	◐	◐		V	Cluster homogeneous traffic
Web Security	Rafique16 [84]	◐	◐	●	M	Infer domain ownership
	Roth20 [89]			●	N	Framing control notification
Other	Stock16 [98]	◐	◐	●	N	Vulnerability disclosure
	Stock18 [97]	◐	◐	●	N	Vulnerability disclosure
	Liu15 [70]	●	●	●	M	Parse WHOIS Records
	Szurdi17 [99]	●		●	M	Group domain owners
	Farooqi17 [47]	●			M	Infer domain ownership

[1] WHOIS fields: R: Registrant, A: Admin, T: Tech, Em: Email address

[2] ●: The paper explicitly describes WHOIS data usage; ◐: No explicit descriptions, but usage can be inferred from context.

[3] WHOIS usage: D: Detection (used for labeling datasets or as features of detection systems), M: Measurement (used for providing measurement results), V: Validation (used for validating results of detection systems), N: Notification (used for reporting vulnerabilities to domain holders)

in response to the GDPR. Table VI characterizes the 35 papers, including the application scenarios and WHOIS fields that are used. Several papers mention WHOIS datasets but do not have a clear description of which fields are used – we infer their WHOIS usage according to the paper context. For the remaining 16 (31%) papers, fields that do not contain personal information are used (e.g., domain creation date and sponsoring registrar). Because they are out of the scope of the GDPR, we consider these works not impacted, and they are characterized in Table VII of Appendix B.

Characterization of WHOIS usage. Here, we provide an in-depth study of the 35 papers relying on the redacted data. As shown in Table VI, the WHOIS database has been used by works on domain security (11 papers), fraud and spam detection (5 papers), cyber-crime analysis (4 papers), privacy protection (3 papers) and HTTPS measurements (3 papers).

We classify WHOIS usage in the papers according to their specific purposes: measurement (22 papers), detection (8 papers), vulnerability notification (4 papers) and result validation (3 papers). Among the redacted WHOIS fields, registrant contact and email address are extensively leveraged, covering 29 and 26 papers. Admin and tech contact are less used, covering 15 papers. In particular, while over 70% GDPR-compliant registrars offer channel to contacting domain holders via web forms or pseudonymized email addresses (per Observation 4), challenges still remain as the scale of notification can be large (e.g., 5K apps in [22] and 24K domains in [97]) and filling the forms automatically is not always feasible. Works using WHOIS for detecting malicious entities or validating results rely on the authentic field values as detection features or ground truth. For measurement studies, the masked fields are heavily used to identify spam campaigns and cluster malicious domain names, thus their accuracy will be impacted as well (e.g., malicious and benign domains could be clustered when their WHOIS records are redacted together by a provider).

Based on Observation 1, the systems developed under those papers have to be re-engineered as the majority of WHOIS providers now follow the GDPR. In addition, as most non-EEA domains are redacted as well (per Observation 5), the impact could be escalated. For attackers exploiting domain registration systems, escaping the detection and tracing becomes much easier, even when they are not EEA citizens.

B. Remediation

The ICANN Temporary Specification proposes a tiered-access framework to be enabled by registrars and registries, in order to allow the usage of WHOIS data under legitimate purposes (e.g., law enforcement and commercial litigation). We find that some providers have already implemented such systems (e.g., `https://tieredaccess.com` by Tucows). However, a recent survey [25] has shown that the issue of data visibility has not been addressed: over 70% access requests have been denied by WHOIS providers under reasons like no court order is shown. The IETF proposed Registration Data Access Protocol (RDAP) [77] to replace the WHOIS protocol in the future, and it is designed to allow specifically authorized people to access private registration data. However, we have not found instructions or links about the authorization process in the existing systems.

In the long term, we believe deploying the tiered-access framework and RDAP is still the right direction, though their operational model should be better designed (e.g., the hurdle to security researchers should be minimized). In the short term, we suggest ICANN refine the requirements listed in the Temporary Specification and suggest a data redaction policy that can balance data utility and privacy. As one example, registration information (including name, email address and street) is very useful for domain clustering. Currently, these fields are usually replaced by fixed values, such as "redacted for privacy" or empty values, making domain clustering less effective. To address this issue, the use of fuzzy hashing [32] could be suggested, which tells the distance between two fields without revealing their original values.

VII. Discussion

Overall, our measurement results show that the GDPR's impact on the Internet domain community is substantial: most WHOIS providers actively redact WHOIS records for compliance. On the other hand, we also find a non-negligible number of partially- or non-compliant providers, with some of them making mistakes during data redaction. Below we list a few recommendations to different stakeholders in hopes of better compliance under the GDPR.

Recommendations. Based on our discussion with registrars, the 1-week window between the ICANN Temporary Specification and GDPR effective date is too short, which leads to excessive data redaction for non-EEA domains. A more efficient format for the discussion between legal authorities and technical communities should be adopted to leave more time for policy execution. This could benefit the stakeholders when new privacy policies like the CCPA are enforced. Secondly, though the language of the ICANN Temporary Specification leaves room for how to redact data, it turns out more confusion is created among WHOIS providers. Implementation flaws have been identified for data redaction. Additionally, as the document leaves flexibility for the data protection scope, some registrars are uncertain about what domains should be protected and therefore sanitize every WHOIS record, though ICANN's intention is to limit the changes to EEA domains only (drawn from our discussions with ICANN staff). We suggest that ICANN make more specific instructions, attach a best practice guidance for technical operators and provide tools for compliance checking instead of relying on complaint reports.

For WHOIS providers, we recommend them to revisit their redaction process and fix errors, e.g., only masking a portion of the registrant's fields. Tools need to be developed to enable periodical inspection and automated error fixing. Auxiliary systems for email forwarding or tiered WHOIS access should be deployed by every provider to ensure legitimate requests (e.g., messages from security researchers) can be fulfilled.

Finally, we also call attention from security researchers and companies who leverage WHOIS data to build applications like malicious domain detection. These applications need to be reassessed or adjusted (e.g., retraining the detection model by removing the redacted fields) to maintain the same level of effectiveness. Alternatively, researchers could push regulators for restricted API access (to accessing data and automatically sending notifications to registrants) and cooperate with large domain registrars for a more uniformed data redaction approach, which balances privacy protection and research needs. An example could be the adoption of fuzzy hashing on the protected fields, instead of using the same redaction string. Moreover, security researchers could also evaluate the influence of different redaction methods.

Online checking tool. To help WHOIS providers gain a clearer view of their current GDPR compliance status of WHOIS data, we design and release an online checking tool (at `https://whoisgdprcompliance.info`). A user representing the provider can view the weekly outlier ratio and a sample of outlying domain names under his/her organization after identity verification. From our discussions with large registrars and ICANN staff, they have shown interest in using the tool.

VIII. Related Work

Works have been published to understand real-world impact of the GDPR. Previous studies focused on web privacy and usable security, as well as compliance checking.

GDPR and web privacy. The expanded territorial scope of the GDPR has pushed many websites to adjust their web privacy policies. By monitoring popular websites of the EU member states and inspecting webpages, [42] reported that many websites have deployed new privacy policies and displayed cookie consent notices. Different types of cookie notices were found to be displayed to visitors according to their country [39]. Recent works measured the impact of the GDPR on the online tracking ecosystem [55], [96] and advertising business [103], [102]. Some of our findings echo with the discoveries on the web, e.g., timely response to the GDPR.

GDPR and usable security. Cookie consent notices become prevalent due to GDPR enforcement and users are required to take actions (e.g., click "Accept Cookies" button) for acknowledgment. How people interact with the notices depends on their implementations, but problematic designs have been identified [104]. When filled with deceptive and misleading language, the consents can be misinterpreted by users [78], [79]. While opt-out options are provided, the practical implementation of web cookies makes it difficult for users to avoid being tracked [90], [50], [49], [27], and some even conflict with the regulations (e.g., mark acceptance even when an explicit opt-out is received from the user) [72]. Overall, these findings call for clearer guidelines for consent notices.

GDPR compliance check. Based on semantic analysis, [80] proposed a framework to analyze legal documents for GDPR compliance and detect violations of privacy norms. Similarly, [45] leveraged knowledge graph to generate rules and regulations mandated for cloud data providers and customers. Automated approaches such as those based on machine learning were proposed to analyze privacy documents [24], [35], [91]. For example, [35] extracted privacy policies based on users' in-context privacy concerns.

GDPR-compliant system design. Some recent works start to rethink the system design in the era of the GDPR or measure the cost of compliance. [92] showed that the performance of a GDPR-compliant database will scale poorly as the volume of personal data increases. Frictions between cloud-scale systems and the GDPR has also been discussed [93]. Those works show building clean-slate GDPR-compliant systems is non-trivial.

IX. Conclusion

In this paper, we report the first systematic and large-scale measurement study on the GDPR compliance process of domain WHOIS data providers, in order to understand the compliance timeline, implementation flaws, scope of protection and collateral damage on security applications. We highlight that the enforcement of the GDPR has brought a profound impact on domain WHOIS services, and identify various flawed implementations of GDPR compliance. We also show that the scope of privacy protection is usually excessive in practice, causing a global impact on the WHOIS system. To quantify the impact on academic research, we conduct a survey study on security papers and find 69% surveyed papers need to use redacted WHOIS information. The results call for a review of current data redaction strategies, and we release an online checking tool to help the stakeholders gain a better view of the compliance status.

Acknowledgments

We thank all anonymous NDSS reviewers for their helpful comments that improved the paper. This work benefits a lot from discussions with Linjian Song and Antonia Nan Chu from Alibaba Cloud, Jianchuan Zhang and Guanghao Li from ICANN, and Dr. Tao Wan. Any views, opinions, findings, recommendations or conclusions contained or expressed herein are those of the authors and do not necessarily reflect official views or policies. This work is supported in part by the National Natural Science Foundation of China (U1836213, U1636204) and the BNRist Network and Software Security Research Program (Grant No. BNR2019TD01004). Baojun Liu is partially supported by the Shuimu Tsinghua Scholar Program.

References

[1] "Directive 95/46/ec of the european parliament and of the council of 24 october 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data," https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:1995:281:TOC, 1995.

[2] "2013 registrar accreditation agreement," https://www.icann.org/en/system/files/files/approved-with-specs-27jun13-en.pdf, 2013.

[3] "Recital 23: Applicable to processors not established in the union if data subjects within the union are targeted," https://gdpr-info.eu/recitals/no-23/, 2016.

[4] "Recital 6: Ensuring a high level of data protection despite the increased exchange of data," https://gdpr-info.eu/recitals/no-6/, 2016.

[5] "Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation)," https://eur-lex.europa.eu/eli/reg/2016/679/oj##ntr4-L_2016119EN.01000101-E0004, 2016.

[6] "Registry agreement," https://newgtlds.icann.org/sites/default/files/agreements/agreement-approved-31jul17-en.pdf, 2017.

[7] "Thick whois transition policy for .com, .net and .jobs," https://www.icann.org/resources/pages/thick-whois-transition-policy-2017-02-01-en, 2017.

[8] "A new era for data protection in the eu: What changes after may 2018," https://ec.europa.eu/info/sites/info/files/data-protection-factsheet-changes_en.pdf, 2018.

[9] "Rdap faqs," https://www.icann.org/resources/pages/rdap-faqs-2018-08-31-en, 2018.

[10] "Ruby whois," https://whoisrb.org/, 2018.

[11] "Temporary specification for gtld registration data," https://www.icann.org/en/system/files/files/gtld-registration-data-temp-spec-17may18-en.pdf, 2018.

[12] "Approved board resolutions regular meeting of the icann board - icann," https://www.icann.org/resources/board-material/resolutions-2019-11-07-en#1.i, 2019.

[13] "California Consumer Privacy Act of 2018," https://iapp.org/resources/article/california-consumer-privacy-act-of-2018/, 2020.

[14] "The chromium projects," https://www.chromium.org/, 2020.

[15] "Free domain privacy and private registration - whoisguard," https://www.namecheap.com/security/whoisguard.aspx, 2020.

[16] "Icann open data platform," https://opendata.icann.org/pages/home-page/, 2020.

[17] "pythonwhois," https://pypi.org/project/pythonwhois/, 2020.

[18] "Registrar ids," https://www.iana.org/assignments/registrar-ids/registrar-ids.xhtml, 2020.

[19] "Root zone database," https://www.iana.org/domains/root/db, 2020.

[20] P. Agten, W. Joosen, F. Piessens, and N. Nikiforakis, "Seven months' worth of mistakes: A longitudinal study of typosquatting abuse," in *Proceedings of the 23rd Network and Distributed System Security Symposium (NDSS 2015)*. Internet Society, 2015.

[21] E. Alowaisheq, P. Wang, S. A. Alrwais, X. Liao, X. Wang, T. Alowaisheq, X. Mi, S. Tang, and B. Liu, "Cracking the wall of confinement: Understanding and analyzing malicious domain takedowns," in *Proceedings of the 26th Network and Distributed System Security Symposium (NDSS 2019)*. Internet Society, 2019.

[22] O. Alrawi, C. Zuo, R. Duan, R. P. Kasturi, Z. Lin, and B. Saltaformaggio, "The betrayal at cloud city: an empirical analysis of cloud-based mobile backends," in *28th USENIX Security Symposium (USENIX Security 19)*, 2019, pp. 551–566.

[23] S. Alrwais, K. Yuan, E. Alowaisheq, Z. Li, and X. Wang, "Understanding the dark side of domain parking," in *23rd USENIX Security Symposium (USENIX Security 14)*, 2014, pp. 207–222.

[24] D. R. Amariles, A. C. Troussel, and R. El Hamdani, "Compliance generation for privacy documents under gdpr: A roadmap for implementing automation and machine learning."

[25] Anti-Phishing Working Group (APWG) and Messaging, Malware and Mobile Anti-Abuse Working Group (M3AAWG), "Icann gdpr and whois users survey," https://docs.apwg.org/reports/ICANN_GDPR_WHOIS_Users_Survey_20181018.pdf, 2018.

[26] Awake Security, "GDPR: Domain security analysis dead end?" https://awakesecurity.com/blog/gdpr-domain-security-analysis/, 2018.

[27] V. Bannihatti Kumar, R. Iyengar, N. Nisal, Y. Feng, H. Habib, P. Story, S. Cherivirala, M. Hagan, L. Cranor, S. Wilson *et al.*, "Finding a choice in a haystack: Automatic extraction of opt-out statements from privacy policy text," in *Proceedings of The Web Conference 2020*, 2020, pp. 1943–1954.

[28] M. A. Bashir, S. Arshad, E. Kirda, W. Robertson, and C. Wilson, "A longitudinal analysis of the ads. txt standard," in *Proceedings of the Internet Measurement Conference*, 2019, pp. 294–307.

[29] M. A. Bashir, S. Arshad, and C. Wilson, "Recommended for you: A first look at content recommendation networks," in *Proceedings of the 2016 Internet Measurement Conference*, 2016, pp. 17–24.

[30] L. Bilge, E. Kirda, C. Kruegel, and M. Balduzzi, "Exposure: Finding malicious domains using passive dns analysis," in *Proceedings of the 18th Network and Distributed System Security Symposium (NDSS 2011)*. Internet Society, 2011.

[31] M. Brandt, T. Dai, A. Klein, H. Shulman, and M. Waidner, "Domain validation++ for mitm-resilient pki," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, 2018, pp. 2060–2076.

[32] F. Breitinger and H. Baier, "A fuzzy hashing approach based on random sequences and hamming distance," in *Proceedings of the Conference on Digital Forensics, Security and Law*. Association of Digital Forensics, Security and Law, 2012, p. 89.

[33] F. Cangialosi, T. Chung, D. Choffnes, D. Levin, B. M. Maggs, A. Mislove, and C. Wilson, "Measurement and analysis of private key sharing in the https ecosystem," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 628–640.

[34] CAUCE North America, "Submission to icann whois team review," https://www.cauce.org/2011/04/submission-to-icann-whois-team-review.html, 2020.

[35] C. Chang, H. Li, Y. Zhang, S. Du, H. Cao, and H. Zhu, "Automated and personalized privacy policy extraction under gdpr consideration," in *International Conference on Wireless Algorithms, Systems, and Applications*. Springer, 2019, pp. 43–54.

[36] Q. A. Chen, E. Osterweil, M. Thomas, and Z. M. Mao, "Mitm attack by name collision: Cause analysis and vulnerability assessment in the new gtld era," in *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2016, pp. 675–690.

[37] N. Christin, S. S. Yanagihara, and K. Kamataki, "Dissecting one click frauds," in *Proceedings of the 17th ACM conference on Computer and communications security*, 2010, pp. 15–26.

[38] A. Cidon, L. Gavish, I. Bleier, N. Korshun, M. Schweighauser, and A. Tsitkin, "High precision detection of business email compromise," in *28th USENIX Security Symposium (USENIX Security 19)*, 2019, pp. 1291–1307.

[39] A. Dabrowski, G. Merzdovnik, J. Ullrich, G. Sendera, and E. Weippl, "Measuring cookies and web privacy in a post-gdpr world," in *In-*

ternational Conference on Passive and Active Network Measurement. Springer, 2019, pp. 258–270.

[40] L. Daigle, “Rfc 3912-whois protocol specification,” *IETF, Sept*, 2004.

[41] J. Dean and S. Ghemawat, “Mapreduce: simplified data processing on large clusters,” *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[42] M. Degeling, C. Utz, C. Lentzsch, H. Hosseini, F. Schaub, and T. Holz, “We value your privacy ... now take some cookies: Measuring the gdpr’s impact on web privacy,” in *Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS 2019)*. Internet Society, 2019.

[43] A. Delignat-Lavaud, M. Abadi, A. Birrell, I. Mironov, T. Wobber, and Y. Xie, “Web pki: Closing the gap between guidelines and practices.” in *Proceedings of the 21st Network and Distributed System Symposium (NDSS 2014)*. Internet Society, 2014.

[44] K. Du, H. Yang, Z. Li, H. Duan, and K. Zhang, “The ever-changing labyrinth: A large-scale analysis of wildcard DNS powered blackhat SEO,” in *25th USENIX Security Symposium (USENIX Security 16)*, 2016, pp. 245–262.

[45] L. Elluri and K. P. Joshi, “A knowledge representation of cloud data controls for eu gdpr compliance,” in *2018 IEEE World Congress on Services (SERVICES)*. IEEE, 2018, pp. 45–46.

[46] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, “A density-based algorithm for discovering clusters in large spatial databases with noise.” in *Kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[47] S. Farooqi, F. Zaffar, N. Leontiadis, and Z. Shafiq, “Measuring and mitigating oauth access token abuse by collusion networks,” in *Proceedings of the 2017 Internet Measurement Conference*, 2017, pp. 355–368.

[48] Farsight Security, “Passive dns historical internet database: Farsight dnsdb,” https://www.farsightsecurity.com/solutions/dnsdb/, 2020.

[49] H. Habib, S. Pearman, J. Wang, Y. Zou, A. Acquisti, L. F. Cranor, N. Sadeh, and F. Schaub, “”it’s a scavenger hunt”: Usability of websites’ opt-out and data deletion choices,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–12.

[50] H. Habib, Y. Zou, A. Jannu, N. Sridhar, C. Swoopes, A. Acquisti, L. F. Cranor, N. Sadeh, and F. Schaub, “An empirical analysis of data deletion and opt-out choices on 150 websites,” in *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, 2019.

[51] T. Halvorson, M. F. Der, I. Foster, S. Savage, L. K. Saul, and G. M. Voelker, “From. academy to. zone: An analysis of the new tld land rush,” in *Proceedings of the 2015 Internet Measurement Conference*, 2015, pp. 381–394.

[52] S. Hao, A. Kantchelian, B. Miller, V. Paxson, and N. Feamster, “Predator: proactive recognition and elimination of domain abuse at time-of-registration,” in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 1568–1579.

[53] S. Hao, M. Thomas, V. Paxson, N. Feamster, C. Kreibich, C. Grier, and S. Hollenbeck, “Understanding the domain registration behavior of spammers,” in *Proceedings of the 2013 Internet measurement conference*, 2013, pp. 63–76.

[54] ICANNWiki, “Whois anti-harvesting techniques — icannwiki,” 2015. [Online]. Available: https://icannwiki.org/index.php?title=Whois_Anti-Harvesting_Techniques&oldid=99496

[55] T. Jakobi, G. Stevens, A. Seufert, and M. Becker, “Webtracking under the new data protection law: Design potentials at the intersection of jurisprudence and HCI,” in *Proceedings of Mensch und Computer 2019*, F. Alt, A. Bulling, and T. Döring, Eds. GI / ACM, 2019, pp. 309–319.

[56] Q. Jenkins, “How has gdpr affected spam?” https://www.spamhaus.org/news/article/775/how-has-gdpr-affected-spam, 2018.

[57] M. T. Khan, X. Huo, Z. Li, and C. Kanich, “Every second counts: Quantifying the negative externalities of cybercrime via typosquatting,” in *2015 IEEE Symposium on Security and Privacy*. IEEE, 2015, pp. 135–150.

[58] A. Kharraz, W. Robertson, and E. Kirda, “Surveylance: automatically detecting online survey scams,” in *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2018, pp. 70–86.

[59] B. Krebs, “Who is afraid of more spams and scams?” https://krebsonsecurity.com/2018/03/who-is-afraid-of-more-spams-and-scams/, 2018.

[60] T. Lauinger, A. S. Buyukkayhan, A. Chaabane, W. Robertson, and E. Kirda, “From deletion to re-registration in zero seconds: Domain registrar behaviour during the drop,” in *Proceedings of the Internet Measurement Conference 2018*, 2018, pp. 322–328.

[61] T. Lauinger, A. Chaabane, A. S. Buyukkayhan, K. Onarlioglu, and W. Robertson, “Game of registrars: An empirical analysis of post-expiration domain name takeovers,” in *26th USENIX Security Symposium (USENIX Security 17)*, 2017, pp. 865–880.

[62] T. Lauinger, K. Onarlioglu, A. Chaabane, W. Robertson, and E. Kirda, “Whois lost in translation: (mis) understanding domain name expiration and re-registration,” in *Proceedings of the 2016 Internet Measurement Conference*, 2016, pp. 247–253.

[63] V. Le Pochat, S. Maroofi, T. Van Goethem, D. Preuveneers, A. Duda, W. Joosen, M. Korczyński *et al.*, “A practical approach for taking down avalanche botnets under real-world constraints,” in *Proceedings of the 27th Network and Distributed System Security Symposium (NDSS 2020)*. Internet Society, 2020.

[64] N. Leontiadis and N. Christin, “Empirically measuring whois misuse,” in *European Symposium on Research in Computer Security*. Springer, 2014, pp. 19–36.

[65] K. Levchenko, A. Pitsillidis, N. Chachra, B. Enright, M. Félegyházi, C. Grier, T. Halvorson, C. Kanich, C. Kreibich, H. Liu *et al.*, “Click trajectories: End-to-end analysis of the spam value chain,” in *2011 ieee symposium on security and privacy*. IEEE, 2011, pp. 431–446.

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

[67] Z. Li and A. Oprea, “Operational security log analytics for enterprise breach detection,” in *2016 IEEE Cybersecurity Development (SecDev)*. IEEE, 2016, pp. 15–22.

[68] D. Liu, S. Hao, and H. Wang, “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*, 2016, pp. 1414–1425.

[69] D. Liu, Z. Li, K. Du, H. Wang, B. Liu, and H. Duan, “Don’t let one rotten apple spoil the whole barrel: Towards automated detection of shadowed domains,” in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 537–552.

[70] S. Liu, I. Foster, S. Savage, G. M. Voelker, and L. K. Saul, “Who is. com? learning to parse whois records,” in *Proceedings of the 2015 Internet Measurement Conference*, 2015, pp. 369–380.

[71] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, “The Stanford CoreNLP natural language processing toolkit,” in *Association for Computational Linguistics (ACL) System Demonstrations*, 2014, pp. 55–60.

[72] C. Matte, N. Bielova, and C. Santos, “Do cookie banners respect my choice?: Measuring legal compliance of banners from iab europe’s transparency and consent framework,” in *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2020, pp. 791–809.

[73] K. McCarthy, “Internet overseer icann loses a third time in whois gdpr legal war,” https://www.theregister.co.uk/2018/08/07/icann_whois_gdpr/, 2018.

[74] D. McGuire, “Ruling on ’.us’ domain raises privacy issues,” https://www.washingtonpost.com/wp-dyn/articles/A7251-2005Mar4.html, 2005.

[75] N. Miramirkhani, O. Starov, and N. Nikiforakis, “Dial one for scam: A large-scale analysis of technical support scams,” in *Proceedings of the 24th Network and Distributed System Security Symposium (NDSS 2017)*. Internet Society, 2017.

[76] H. Mohajeri Moghaddam, G. Acar, B. Burgess, A. Mathur, D. Y. Huang, N. Feamster, E. W. Felten, P. Mittal, and A. Narayanan, “Watching you watch: The tracking ecosystem of over-the-top tv streaming devices,” in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019, pp. 131–147.

[77] A. Newton and S. Hollenbeck, “Registration data access protocol (rdap) query format,” RFC 7482, February 2015, http://www. rfc-editor. org/info/rfc7482, Tech. Rep., 2015.

[78] M. Nouwens, I. Liccardi, M. Veale, D. Karger, and L. Kagal, "Dark patterns after the gdpr: Scraping consent pop-ups and demonstrating their influence," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–13.

[79] E. Okoyomon, N. Samarin, P. Wijesekera, A. Elazari Bar On, N. Vallina-Rodriguez, I. Reyes, Á. Feal, and S. Egelman, "On the ridiculousness of notice and consent: Contradictions in app privacy policies," in *The Workshop on Technology and Consumer Protection (ConPro'19)*, 2019.

[80] M. Palmirani and G. Governatori, "Modelling legal knowledge for gdpr compliance checking." in *JURIX*, 2018, pp. 101–110.

[81] V. Paxson, M. Christodorescu, M. Javed, J. Rao, R. Sailer, D. L. Schales, M. Stoecklin, K. Thomas, W. Venema, and N. Weaver, "Practical comprehensive bounds on surreptitious communication over DNS," in *22nd USENIX Security Symposium (USENIX Security 13)*, 2013, pp. 17–32.

[82] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[83] D. Plohmann, K. Yakdan, M. Klatt, J. Bader, and E. Gerhards-Padilla, "A comprehensive measurement study of domain generating malware," in *25th USENIX Security Symposium (USENIX Security 16)*, 2016, pp. 263–278.

[84] M. Z. Rafique, T. Van Goethem, W. Joosen, C. Huygens, and N. Nikiforakis, "It's free for a reason: Exploring the ecosystem of free live streaming services," in *Proceedings of the 23rd Network and Distributed System Security Symposium (NDSS 2016)*. Internet Society, 2016, pp. 1–15.

[85] M. S. Rahman, T.-K. Huang, H. V. Madhyastha, and M. Faloutsos, "Efficient and scalable socware detection in online social networks," in *21st USENIX Security Symposium (USENIX Security 12)*, 2012, pp. 663–678.

[86] B. Reaves, N. Scaife, D. Tian, L. Blue, P. Traynor, and K. R. Butler, "Sending out an sms: Characterizing the security of the sms ecosystem with public gateways," in *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2016, pp. 339–356.

[87] J. Ren, M. Lindorfer, D. J. Dubois, A. Rao, D. Choffnes, and N. Vallina-Rodriguez, "A longitudinal study of pii leaks across android app versions," in *Proceedings of the 25th Network and Distributed System Security Symposium (NDSS 2018)*, 2018.

[88] R. Roberts, Y. Goldschlag, R. Walter, T. Chung, A. Mislove, and D. Levin, "You are who you appear to be: A longitudinal study of domain impersonation in tls certificates," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019, pp. 2489–2504.

[89] S. Roth, T. Barron, S. Calzavara, N. Nikiforakis, and B. Stock, "Complex security policy? A longitudinal analysis of deployed content security policies," in *Proceedings of the 27th Network and Distributed System Security Symposium (NDSS 2020)*. Internet Society, 2020.

[90] I. Sanchez-Rola, M. Dell'Amico, P. Kotzias, D. Balzarotti, L. Bilge, P.-A. Vervier, and I. Santos, "Can i opt out yet? gdpr and the global illusion of cookie control," in *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*, 2019, pp. 340–351.

[91] D. Sarne, J. Schler, A. Singer, A. Sela, and I. Bar Siman Tov, "Unsupervised topic extraction from privacy policies," in *Companion Proceedings of The 2019 World Wide Web Conference*, 2019, pp. 563–568.

[92] S. Shastri, V. Banakar, M. Wasserman, A. Kumar, and V. Chidambaram, "Understanding and benchmarking the impact of gdpr on database systems," *Proceedings of the VLDB Endowment*, vol. 13, no. 7, pp. 1064–1077, 2020.

[93] S. Shastri, M. Wasserman, and V. Chidambaram, "GDPR anti-patterns: How design and operation of modern cloud-scale systems conflict with GDPR," *CoRR*, vol. abs/1911.00498, 2019. [Online]. Available: http://arxiv.org/abs/1911.00498

[94] A. Singhal, C. Buckley, and M. Mitra, "Pivoted document length normalization," in *ACM SIGIR Forum*, vol. 51, no. 2. ACM New York, NY, USA, 2017, pp. 176–184.

[95] S. Sivakorn, K. Jee, Y. Sun, L. Kort-Parn, Z. Li, C. Lumezanu, Z. Wu, L.-A. Tang, and D. Li, "Countering malicious processes with process-dns association." in *Proceedings of the 26th Network and Distributed Systems Security Symposium (NDSS 2019)*. Internet Society, 2019.

[96] K. Solomos, P. Ilia, S. Ioannidis, and N. Kourtellis, "Clash of the trackers: Measuring the evolution of the online tracking ecosystem," *CoRR*, vol. abs/1907.12860, 2019. [Online]. Available: http://arxiv.org/abs/1907.12860

[97] B. Stock, G. Pellegrino, F. Li, M. Backes, and C. Rossow, "Didn't you hear me? - towards more successful web vulnerability notifications," in *Proceedings of the 25th Network and Distributed System Security Symposium (NDSS 2018)*. Internet Society, 2018.

[98] B. Stock, G. Pellegrino, C. Rossow, M. Johns, and M. Backes, "Hey, you have a problem: On the feasibility of large-scale web vulnerability notification," in *25th USENIX Security Symposium (USENIX Security 16)*, 2016, pp. 1015–1032.

[99] J. Szurdi and N. Christin, "Email typosquatting," in *Proceedings of the 2017 Internet Measurement Conference*, 2017, pp. 419–431.

[100] J. Szurdi, B. Kocso, G. Cseh, J. Spring, M. Felegyhazi, and C. Kanich, "The long "taile" of typosquatting domain names," in *23rd USENIX Security Symposium (USENIX Security 14)*, 2014, pp. 191–206.

[101] K. Tian, S. T. Jan, H. Hu, D. Yao, and G. Wang, "Needle in a haystack: Tracking down elite phishing domains in the wild," in *Proceedings of the Internet Measurement Conference 2018*, 2018, pp. 429–442.

[102] T. Urban, D. Tatang, M. Degeling, and T. Holz, "A study on subject data access in online advertising after the gdpr," *Data Privacy Management, Cryptocurrencies and Blockchain Technology*, p. 61.

[103] T. Urban, D. Tatang, M. Degeling, T. Holz, and N. Pohlmann, "Measuring the impact of the gdpr on data sharing in ad networks," in *Proceedings of the 15th ACM ASIA Conference on Computer and Communications Security, ASIA CCS*, vol. 20, 2020.

[104] C. Utz, M. Degeling, S. Fahl, F. Schaub, and T. Holz, "(un) informed consent: Studying gdpr consent notices in the field," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019, pp. 973–990.

[105] P. Vallina, Á. Feal, J. Gamba, N. Vallina-Rodriguez, and A. F. Anta, "Tales from the porn: A comprehensive privacy analysis of the web porn ecosystem," in *Proceedings of the Internet Measurement Conference*, 2019, pp. 245–258.

[106] T. van Ede, R. Bortolameotti, A. Continella, J. Ren, D. J. Dubois, M. Lindorfer, D. Choffnes, M. van Steen, and A. Peter, "Flowprint: Semi-supervised mobile-app fingerprinting on encrypted network traffic," in *Proceedings of the 27th Network and Distributed System Security Symposium (NDSS 2020)*. Internet Society, 2020.

[107] T. Vissers, T. Barron, T. Van Goethem, W. Joosen, and N. Nikiforakis, "The wolf of name street: Hijacking domains through their nameservers," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 957–970.

[108] T. Vissers, W. Joosen, and N. Nikiforakis, "Parking sensors: Analyzing and detecting parked domains," in *Proceedings of the 22nd Network and Distributed System Security Symposium (NDSS 2015)*. Internet Society, 2015, pp. 53–53.

[109] D. Wang, S. Savage, and G. M. Voelker, "Juice: A longitudinal study of an seo campaign," in *Proceedings of the 20th Nework and Distributed Systems Security Symposium (NDSS 2013)*. Internet Society, 2013.

[110] Wikipedia contributors, "Message transfer agent — Wikipedia, the free encyclopedia," https://en.wikipedia.org/w/index.php?title=Message_transfer_agent&oldid=936498407, 2020.

[111] S. Yadav, A. K. K. Reddy, A. N. Reddy, and S. Ranjan, "Detecting algorithmically generated malicious domain names," in *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, 2010, pp. 48–61.

[112] H. Yang, X. Ma, K. Du, Z. Li, H. Duan, X. Su, G. Liu, Z. Geng, and J. Wu, "How to learn klingon without a dictionary: Detection and measurement of black keywords used by the underground economy," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 751–769.

[113] S. Zimmeck, J. S. Li, H. Kim, S. M. Bellovin, and T. Jebara, "A privacy analysis of cross-device tracking," in *26th USENIX Security Symposium (USENIX Security 17)*, 2017, pp. 1391–1408.

TABLE VII: Security literature that use non-redacted WHOIS data

Category	Reference	WHOIS Fields	WHOIS Usage	Details
Domain Security	Yadav10 [111]	No specific descriptions	Validation	Check suspicious samples
	Bilge11 [30]	No specific descriptions	Detection	Collect benign datasets
	Szurdi14 [100]	Registrar, Creation Date	Detection	Features for detection
	Agten15 [20]	No specific descriptions	Detection, Measurement	Identify abuse type, Analyze malicious behaviors
	Liu16 [68]	Registrar, Updated Date, Creation Date, Expiration Date	Detection, Measurement	Check expired domains, Threat analysis
	Hao16 [52]	Registrar, Nameserver IP and AS, Creation Date, Expiration Date	Detection	Features for detection
	Lauinger16 [62]	Creation Date, Updated Date, Expiration Date	Detection, Measurement	Discover expiring domain, Analyze re-registration
	Lauinger17 [61]	Registrar, Creation Date, Updated Date, Expiration Date	Measurement	Check domain registration status
	Lauinger18 [60]	Registrar, Creation Date, Expiration Date	Measurement	Track domain registration status
Spam & Phishing	Levchenko11 [65]	Registrar, Nameserver IP	Measurement	Analyze spam infrastructure
	Hao13 [53]	Registrar, historical WHOIS information	Measurement	Analyze spam registration behavior
	Tian18 [101]	Registrar, Creation Date	Measurement	Analyze phishing registration behavior
Online Social Networks	Rahman12 [85]	No specific descriptions	Measurement	Check detected URLs
	Bashir16 [29]	Creation Date	Measurement	Measure age of landing domains
Privacy	Moghaddam19 [76]	No specific descriptions	Measurement	Remove false positives
Email Security	Cidon19 [38]	Creation Date	Detection	Features for detection

Appendix A
Registrar Domain Share

As required by the RA [6], gTLD registries submit monthly reports about the domain names they sponsor. The Per-Registrar Transactions Reports record the number of total domains sponsored by each registrar ID. The reports are released on the ICANN Open Data Platform [16] and we download the latest version available which was released in Nov 2019. For each registrar, we calculate the percentage of its sponsored domains to indicate its share of the domain business. Table VIII shows the domain share of the top 25 registrars.

TABLE VIII: Share of registered domains of the top 25 registrars by ID (Nov 2019).

Rank	ID	Registrar Name	# Total Domains	Share
1	146	GoDaddy.com, LLC	61,645,127	29.09%
2	69	Tucows Domains Inc.	9,926,177	4.68%
3	1068	NameCheap, Inc.	9,454,269	4.46%
4	2	Network Solutions, LLC	7,011,438	3.31%
5	420	Alibaba Cloud Computing (Beijing) Co., Ltd.	6,824,144	3.22%
6	48	eNom, LLC	5,590,700	2.64%
7	49	GMO Internet, Inc. d/b/a Onamae.com	5,400,764	2.55%
8	1599	Alibaba Cloud Computing Ltd. d/b/a HiChina (www.net.cn)	5,381,119	2.54%
9	120	Xin Net Technology Corporation	4,966,779	2.34%
10	83	1&1 IONOS SE	4,925,377	2.32%
11	303	PDR Ltd. d/b/a PublicDomainRegistry.com	4,572,228	2.16%
12	895	Google LLC	4,111,495	1.94%
13	1556	Chengdu West Dimension Digital Technology Co., Ltd.	3,369,930	1.59%
14	1479	NameSilo, LLC	3,368,280	1.59%
15	440	Wild West Domains, LLC	2,731,433	1.29%
16	1154	FastDomain Inc.	2,323,022	1.10%
17	433	OVH sas	2,226,535	1.05%
18	472	Dynadot, LLC	1,973,675	0.93%
19	625	Name.com, Inc.	1,963,699	0.93%
20	1915	West263 International Limited	1,870,561	0.88%
21	9	Register.com, Inc.	1,740,346	0.82%
22	886	Domain.com, LLC	1,738,049	0.82%
23	460	Web Commerce Communications Limited dba WebNic.cc	1,679,382	0.79%
24	269	Key-Systems GmbH	1,391,870	0.66%
25	299	CSC Corporate Domains, Inc.	1,377,711	0.65%

Appendix B
Security Literature that Use Non-Redacted WHOIS Fields

In Table VII we summarize 16 security papers using WHOIS information which are considered as not impacted by the GDPR. For detection and measurement purposes, the papers use fields such as dates (e.g., domain creation and expiration dates) and registrar identity information (e.g., registrar ID). These fields do not contain personal data and are thus not required to be redacted according to the ICANN Temporary Specification.