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 1911

UCI University of California, Irvine

奇安信 QI-ANXIN

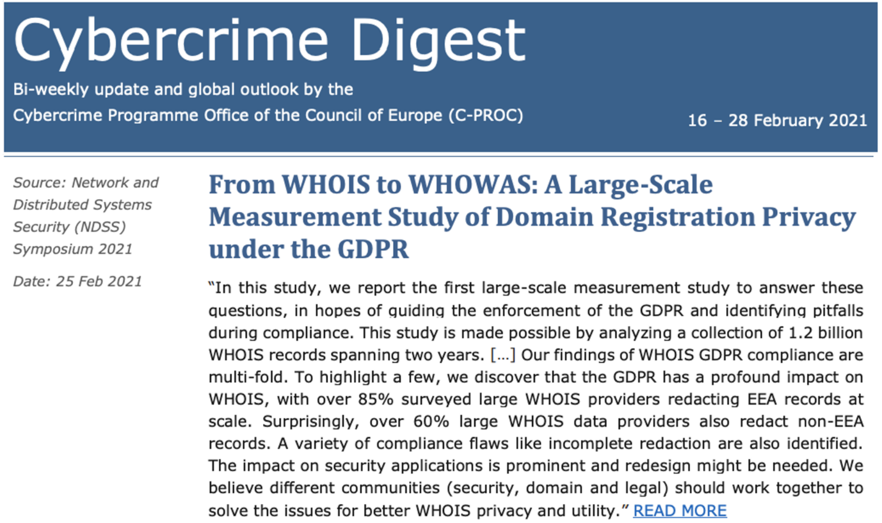
Netlab 360.com

UT DALLAS The University of Texas at Dallas

FUDAN UNIVERSITY 1905

Media Reports

Cybercrime Programme Office of the Council of Europe



Cybercrime Digest

Bi-weekly update and global outlook by the
Cybercrime Programme Office of the Council of Europe (C-PROC)

16 – 28 February 2021

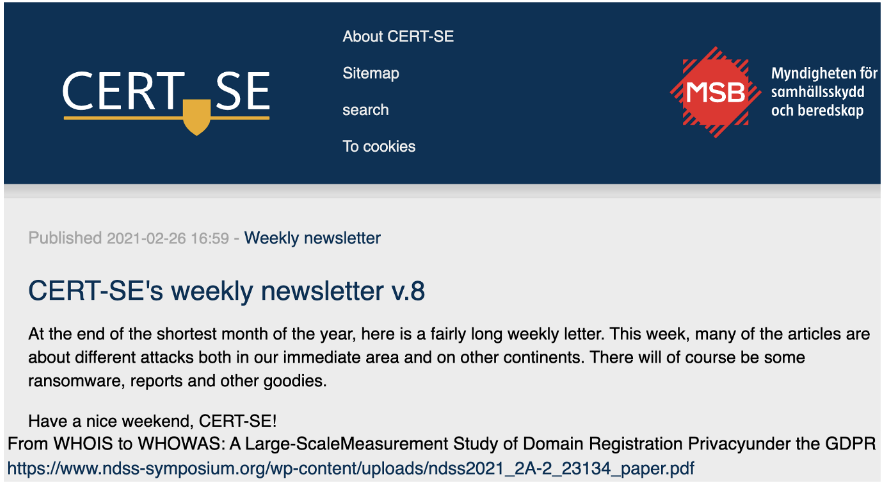
Source: Network and Distributed Systems Security (NDSS) Symposium 2021

Date: 25 Feb 2021

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

"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. […] 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." READ MORE

CERT-SE (Computer Security Incident Response Team of Sweden)



Published 2021-02-26 16:59 - Weekly newsletter

CERT-SE's weekly newsletter v.8

At the end of the shortest month of the year, here is a fairly long weekly letter. This week, many of the articles are about different attacks both in our immediate area and on other continents. There will of course be some ransomware, reports and other goodies.

Have a nice weekend, CERT-SE!

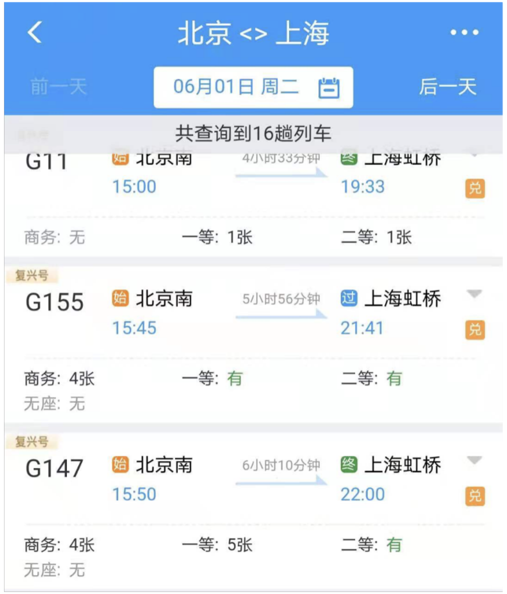
From WHOIS to WHOWAS: A Large-ScaleMeasurement Study of Domain Registration Privacyunder the GDPR
https://www.ndss-symposium.org/wp-content/uploads/ndss2021_2A-2_23134_paper.pdf

When Systems Go Real-Name...

Defeats abusive acts effectively

Cellular networks

Transportation



北京 <> 上海

前一天 06月01日 周二 后一天

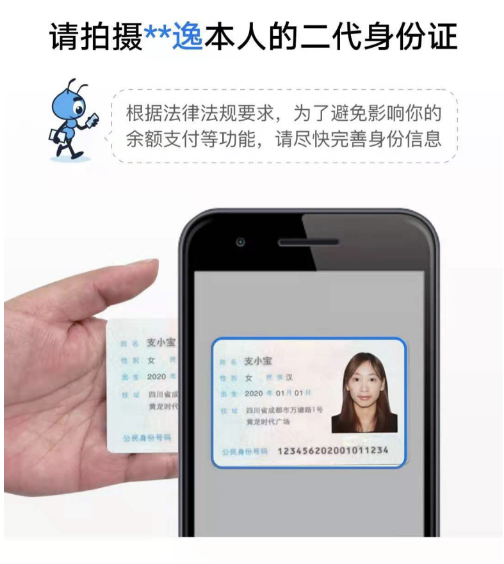
共查询到16趟列车

G11 始 北京南 4小时33分钟 终 上海虹桥
15:00 19:33 兑
商务：无 一等：1张 二等：1张

复兴号
G155 始 北京南 5小时56分钟 过 上海虹桥
15:45 21:41 兑
商务：4张 一等：有 二等：有
无座：无

复兴号
G147 始 北京南 6小时10分钟 终 上海虹桥
15:50 22:00 兑
商务：4张 一等：5张 二等：有
无座：无

Online activity

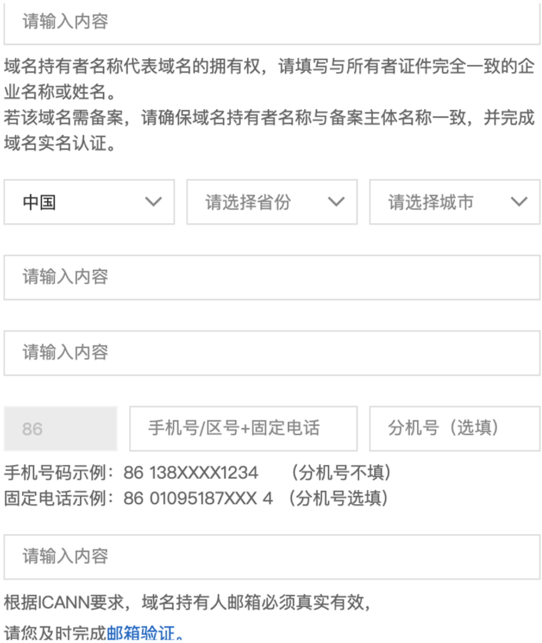


请拍摄**逸本人的二代身份证

根据法律法规要求，为了避免影响你的余额支付等功能，请尽快完善身份信息

Domain Registration Goes Real-Name, Too

Supported by ICANN and government regulations



Registrant Name

请输入内容

域名持有者名称代表域名的拥有权，请填写与所有者证件完全一致的企业名称或姓名。
若该域名需备案，请确保域名持有者名称与备案主体名称一致，并完成域名实名认证。

Postal Address & Code

中国 | 请选择省份 | 请选择城市

请输入内容

请输入内容

Phone

86 | 手机号/区号+固定电话 | 分机号（选填）

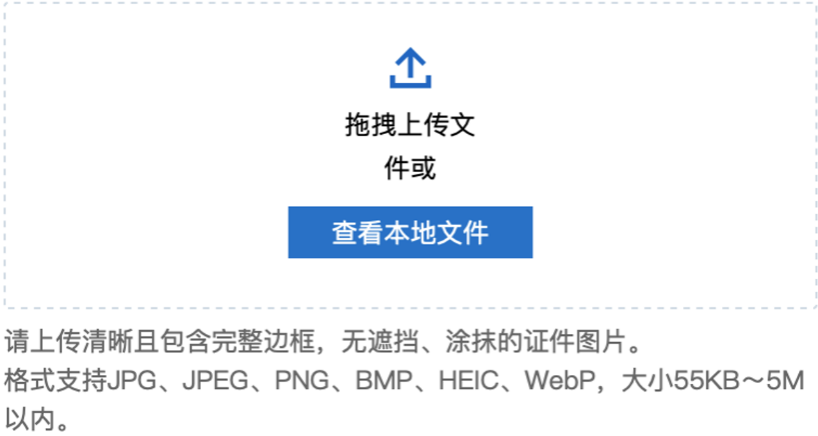
手机号码示例：86 138XXXX1234 （分机号不填）
固定电话示例：86 01095187XXX 4 （分机号选填）

Email

请输入内容

根据ICANN要求，域名持有人邮箱必须真实有效，
请您及时完成邮箱验证。

ID card & Passport verification



请上传清晰且包含完整边框，无遮挡、涂抹的证件图片。
格式支持JPG、JPEG、PNG、BMP、HEIC、WebP，大小55KB～5M以内。

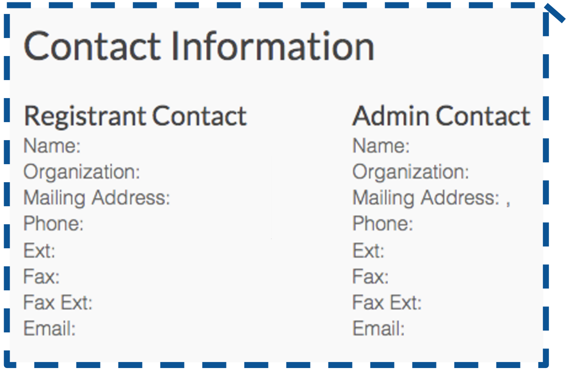
(Domain registration data required by AliYun)

WHOIS: Real-Name for Domain Registration

Personal data of domain holders are *collected*

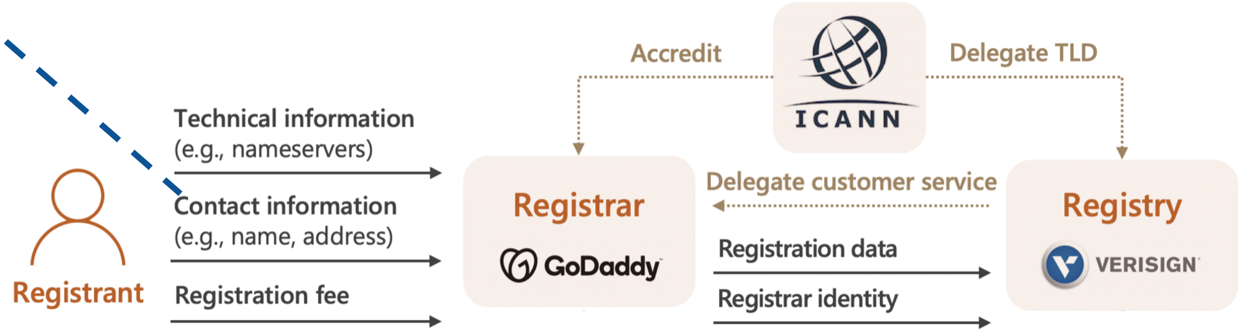
Names, addresses, phone numbers and emails

Stored by registrars and registries (WHOIS *providers*)



Contact Information

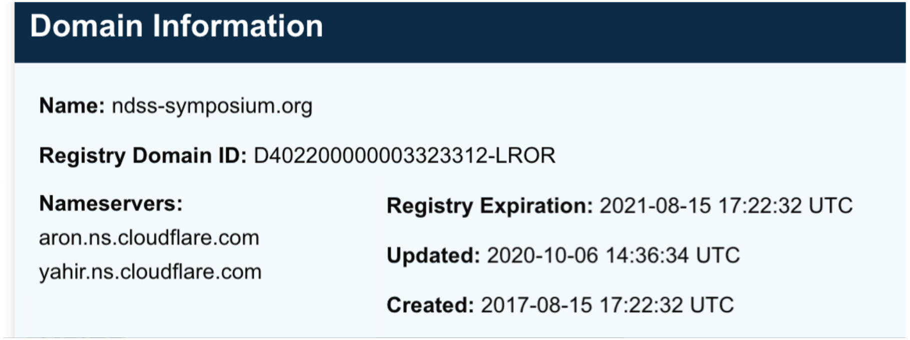
Registrant Contact	Admin Contact
Name:	Name:
Organization:	Organization:
Mailing Address:	Mailing Address: ,
Phone:	Phone:
Ext:	Ext:
Fax:	Fax:
Fax Ext:	Fax Ext:
Email:	Email:



WHOIS: Real-Name for Domain Registration

Personal data of domain holders are *published*

- Query-based access via <u>*WHOIS protocol*</u>
- Web-based interface / WHOIS server (TCP port 43)
- <u>*WHOIS query is open and free to everyone*</u>



Domain Information

Name: ndss-symposium.org

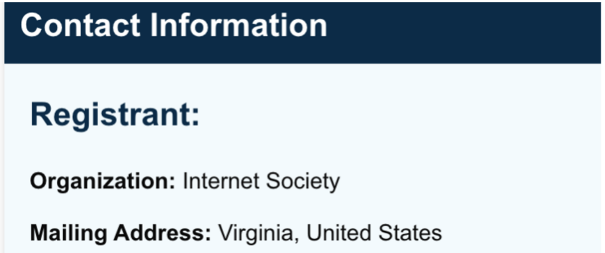
Registry Domain ID: D402200000003323312-LROR

Nameservers:
aron.ns.cloudflare.com
yahir.ns.cloudflare.com

Registry Expiration: 2021-08-15 17:22:32 UTC

Updated: 2020-10-06 14:36:34 UTC

Created: 2017-08-15 17:22:32 UTC



Contact Information

Registrant:

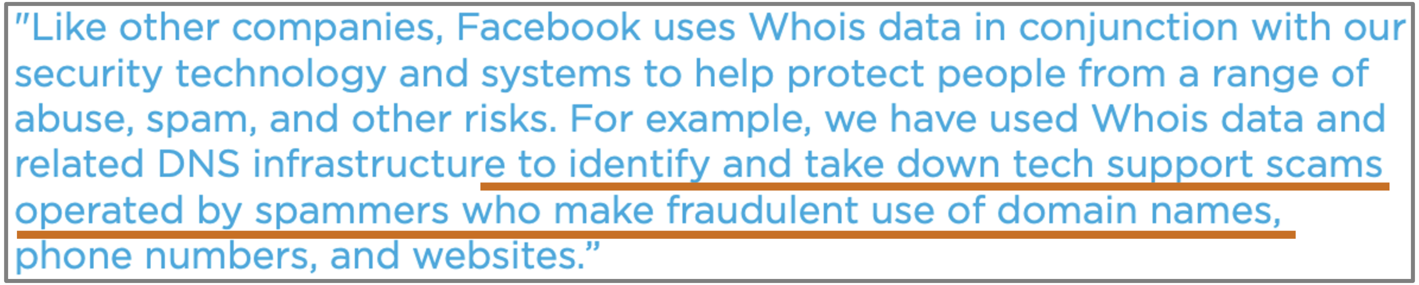
Organization: Internet Society

Mailing Address: Virginia, United States

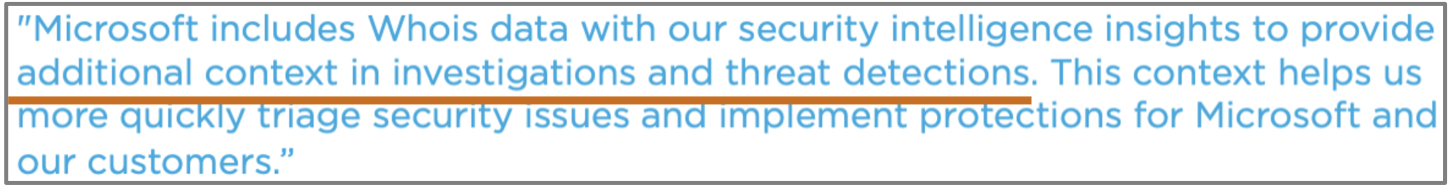
(Domain registration data of **ndss-symposium.org**
acquired from lookup.icann.org on Jan 31, 2021)

Security Feeds on WHOIS, Heavily

Spam detection, domain takedown, vulnerability notification...



"Like other companies, Facebook uses Whois data in conjunction with our security technology and systems to help protect people from a range of abuse, spam, and other risks. For example, we have used Whois data and related DNS infrastructure to identify and take down tech support scams operated by spammers who make fraudulent use of domain names, phone numbers, and websites."



"Microsoft includes Whois data with our security intelligence insights to provide additional context in investigations and threat detections. This context helps us more quickly triage security issues and implement protections for Microsoft and our customers."

* https://www.icann.org/en/system/files/files/gdpr-white-paper-domain-tools-icann-proposed-compliance-models-18feb18-en.pdf

Sounds good, right?

Until…

General Data Protection Regulation

A high-level framework about protecting personal data

- Personal data: information of identifying/identifiable natural person
- Protects personal data *processing* (storage, disclosure, ...)

General Data Protection Regulation

A high-level framework about protecting personal data

Personal data: information of identifying/identifiable natural person

Protects personal data *processing* (storage, disclosure, ...)

Expanded territorial scope

Applies to processing of personal data of subjects in the EU

Regardless of where the processing takes place

Profound impact on Internet applications

Website cookies, online ads, privacy notices, ...

When WHOIS Meets GDPR

"WHOIS" becomes *"WHOWAS"*

Releasing personal data in WHOIS shall be consented

Guidelines published by ICANN on May 17, 2018

"Temporary Specification for gTLD Registration Data"* (TempSpec)

Applies to all gTLD registries and registrars

Collection of registration data

Is maintained.

Personal data is still collected at domain registration.

Access to registration data

Is restricted.

Tiered/layered access under legitimate purposes.

* https://www.icann.org/en/system/files/files/gtld-registration-data-temp-spec-17may18-en.pdf

When WHOIS Meets GDPR

WHOIS publishing requirements of ICANN TempSpec

Replacing personal data with *redacted/anonymized* values

Providers decide the scope of data to be protected.

Registration Data Fields	Data Subjects	Data Publishing Requirements
Name, Street, City, Postal Code, Phone, Fax	Registrant, Admin, Tech	1) Provide a **redacted value** (“*substantially similar*” to “redacted for privacy”), or 2) Provide an **empty value**, or do not provide the fields
Organization, State/Province, Country	Admin, Tech	
Email Address	Registrant, Admin, Tech	Provide an **anonymized email address** or **web form** enabling communication with data subject

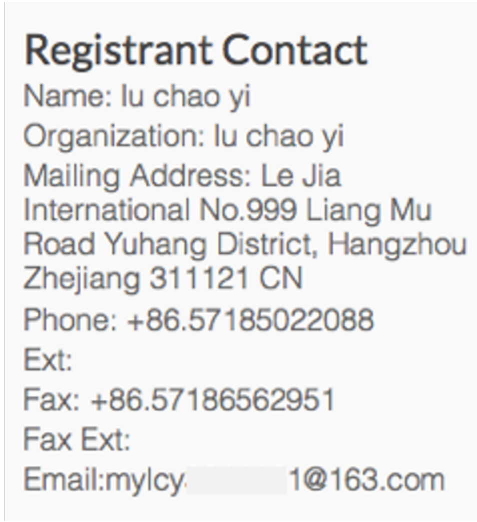
* https://www.icann.org/en/system/files/files/gtld-registration-data-temp-spec-17may18-en.pdf

When WHOIS Meets GDPR

WHOIS publishing requirements of ICANN TempSpec

Replacing personal data with *redacted/anonymized* values

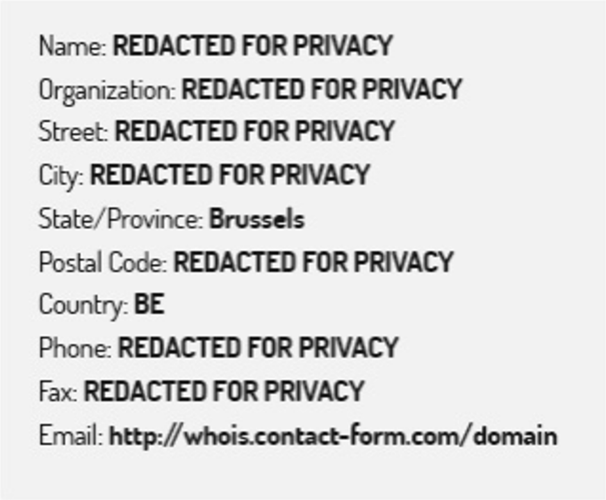
Providers decide the scope of data to be protected.



Registrant Contact

Name: lu chao yi
Organization: lu chao yi
Mailing Address: Le Jia International No.999 Liang Mu Road Yuhang District, Hangzhou Zhejiang 311121 CN
Phone: +86.57185022088
Ext:
Fax: +86.57186562951
Fax Ext:
Email:mylcy[redacted]1@163.com

Not protected



Name: **REDACTED FOR PRIVACY**
Organization: **REDACTED FOR PRIVACY**
Street: **REDACTED FOR PRIVACY**
City: **REDACTED FOR PRIVACY**
State/Province: **Brussels**
Postal Code: **REDACTED FOR PRIVACY**
Country: **BE**
Phone: **REDACTED FOR PRIVACY**
Fax: **REDACTED FOR PRIVACY**
Email: **http://whois.contact-form.com/domain**

Redacted

Research Questions

Data Publishing Changes of WHOIS Providers

Are providers compliant to the TempSpec?

How do they redact WHOIS data?

Are there any compliance flaws?

What is the scope of protected domains?

Security Impact of WHOIS Data Loss

How many security works rely on WHOIS?

Do they use redacted WHOIS data?

What are the security systems used for?

How to remediate the loss of WHOIS?

Part I-A:

Data Publishing Changes of WHOIS Providers

(Methodology)

Methodology: Overview

Data-driven measurement study

Latitudinal view: covering a wide range of WHOIS providers

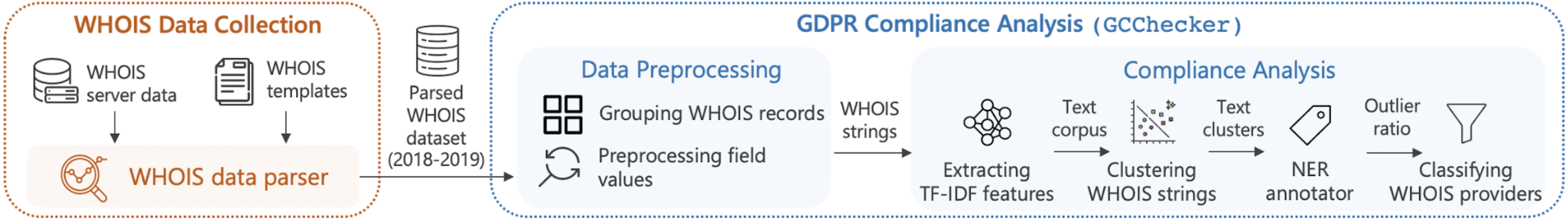
Longitudinal view: covering dates before/after GDPR went effective

A. WHOIS data collection

2-year parsed WHOIS data

B. Compliance Analysis (*GCChecker*)

Identify protected/redacted records and give compliance rankings



Methodology: WHOIS Data Collection

Challenge: WHOIS ecosystem is fragmented

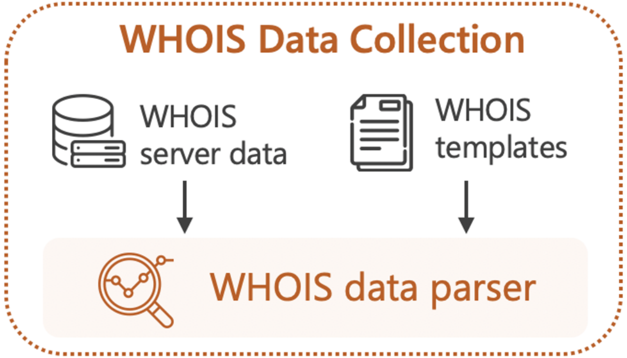
Hundreds of providers maintain WHOIS servers

Format of WHOIS data is *inconsistent*

Solution: parsed historical WHOIS dataset from industrial partner

Collects WHOIS of domains observed in its passive DNS

Parsed by *manually-generated templates*



Methodology: WHOIS Data Collection

Overview of WHOIS dataset (Jan 2018 ~ Dec 2019)

12% EEA domains; 13% domains older than 10 years

Collected from port 43 of WHOIS servers (not from web WHOIS tools)

Year	Count of				Creation Date		Registrant Region	
	Record	Domain	Region	TLD	~ ’09	’10 ~ ’19	EEA	Non-EEA
2018	659M	211M	218	758	15.7%	84.3%	12.9%	87.1%
2019	583M	215M	218	754	14.5%	85.5%	12.4%	87.6%
All	**1.24B**	**267M**	**219**	**783**	**13.4%**	**86.6%**	**12.2%**	**87.8%**

Methodology: Compliance Analysis

Challenge: different wording/language for redaction

TempSpec do not enforce the use of *“redacted for privacy”*

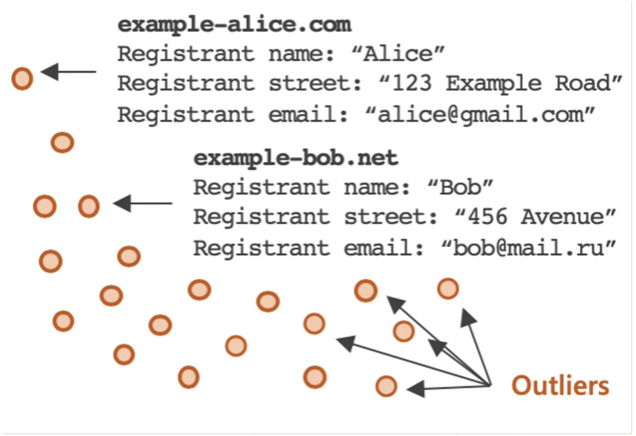
Methodology: Compliance Analysis

Challenge: different wording/language for redaction

TempSpec do not enforce the use of *"redacted for privacy"*

Solution: unsupervised clustering of WHOIS record groups

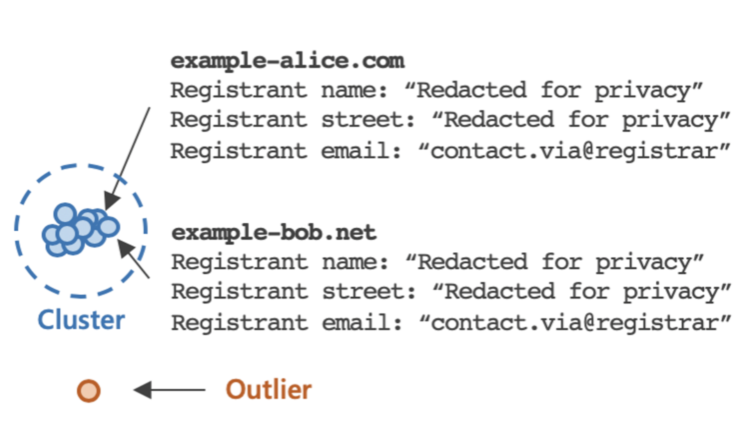
Replace records at scale → High textual similarity → Clusters → Few *Outliers*



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"

Not compliant, %outlier is high



example-alice.com
Registrant name: "Redacted for privacy"
Registrant street: "Redacted for privacy"
Registrant email: "contact.via@registrar"

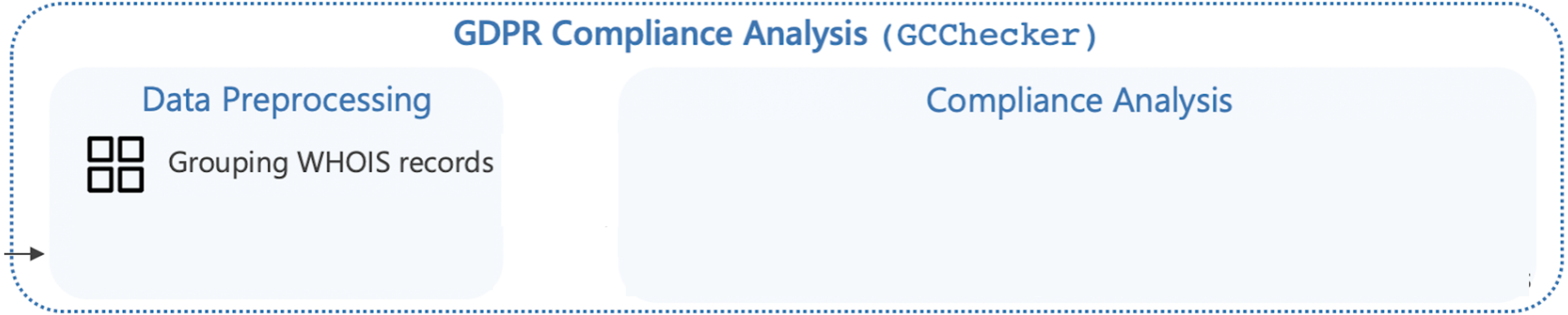
example-bob.net
Registrant name: "Redacted for privacy"
Registrant street: "Redacted for privacy"
Registrant email: "contact.via@registrar"

Compliant, %outlier is low

Methodology: Compliance Analysis

Design of *GCChecker*

Grouping WHOIS records: *(provider, registrant_region, data_subject, week)*

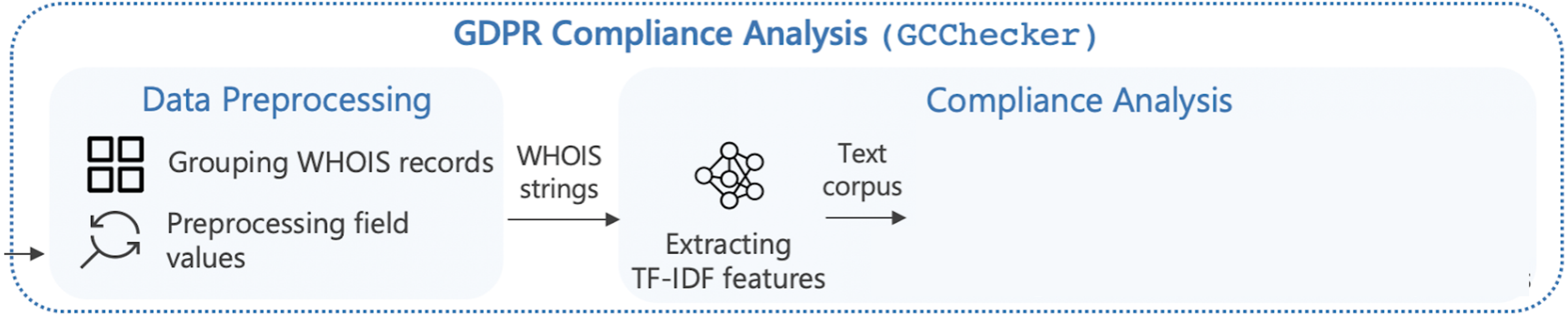


Methodology: Compliance Analysis

Design of *GCChecker*

Grouping WHOIS records: *(provider, registrant_region, data_subject, week)*

Preprocessing: normalize values, extract *TF-IDF features*


GDPR Compliance Analysis (GCChecker)
Data Preprocessing
Grouping WHOIS records
Preprocessing field values
WHOIS strings
Compliance Analysis
Extracting TF-IDF features
Text corpus

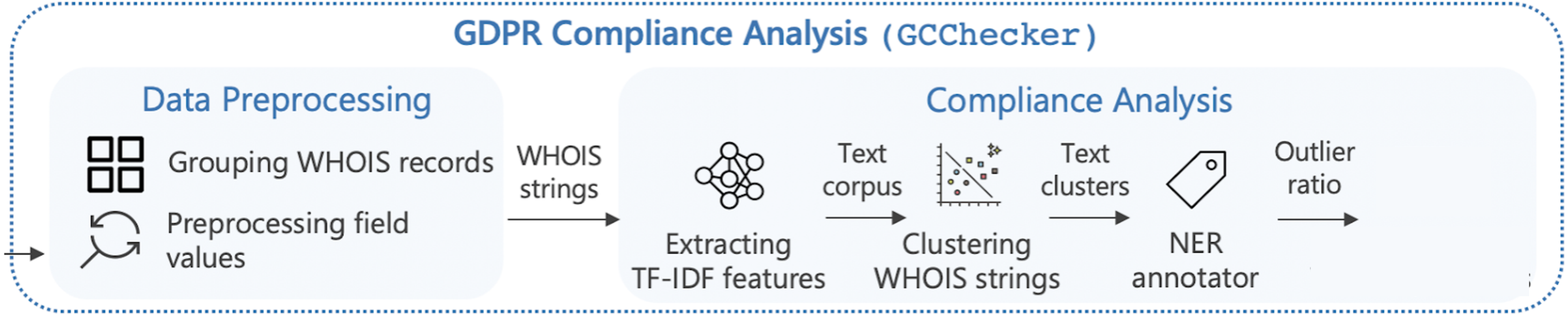
Methodology: Compliance Analysis

Design of *GCChecker*

Grouping WHOIS records: *(provider, registrant_region, data_subject, week)*

Preprocessing: normalize values, extract *TF-IDF features*

Clustering: DBSCAN finds *outliers*, NER refines clusters



Methodology: Compliance Analysis

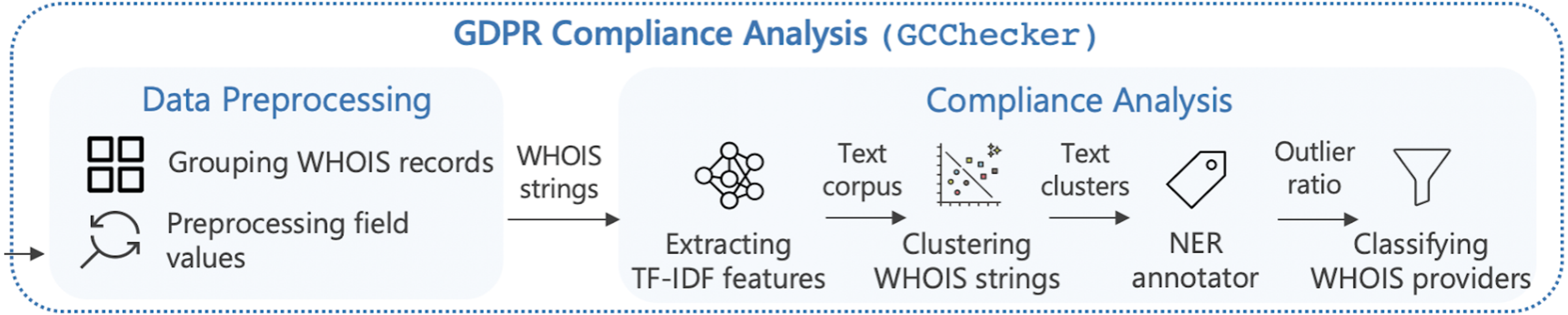
Design of *GCChecker*

Grouping WHOIS records: *(provider, registrant_region, data_subject, week)*

Preprocessing: normalize values, extract *TF-IDF features*

Clustering: DBSCAN finds *outliers*, NER refines clusters

Provider classification: rank from on weekly outlier ratios



Part I-B:

Data Publishing Changes of WHOIS Providers

(Results of 143 large providers)

Scale of WHOIS Data Redaction

Over 85% large WHOIS providers are fully-compliant

Large: as of *EEA WHOIS records* collected

Registrars: 73 / 89 (total domain share > 54%)

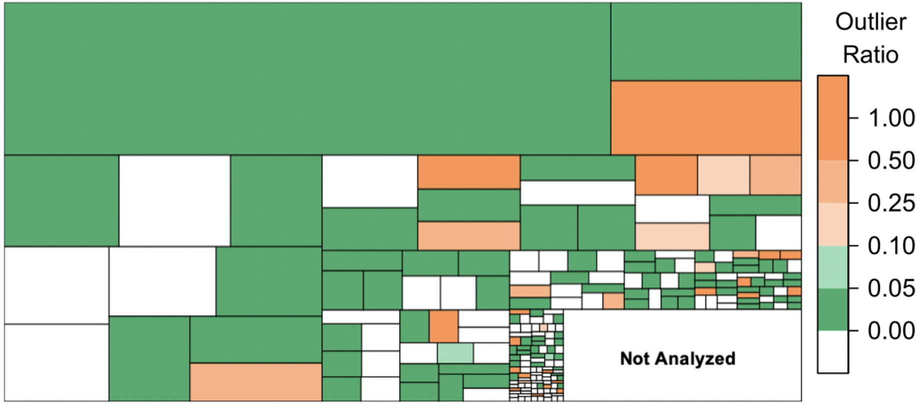
Registries: 51 / 54

Flawed implementations

Missing protection of addresses

Only protecting email addresses

Others...



WHOIS compliance of EEA records from registrars (corresponding with their domain share)

Timeline of WHOIS Data Redaction

Over 80% fully-compliant providers completed in time

100 / 124 completed before May 25, 2018

Timeline of WHOIS Data Redaction

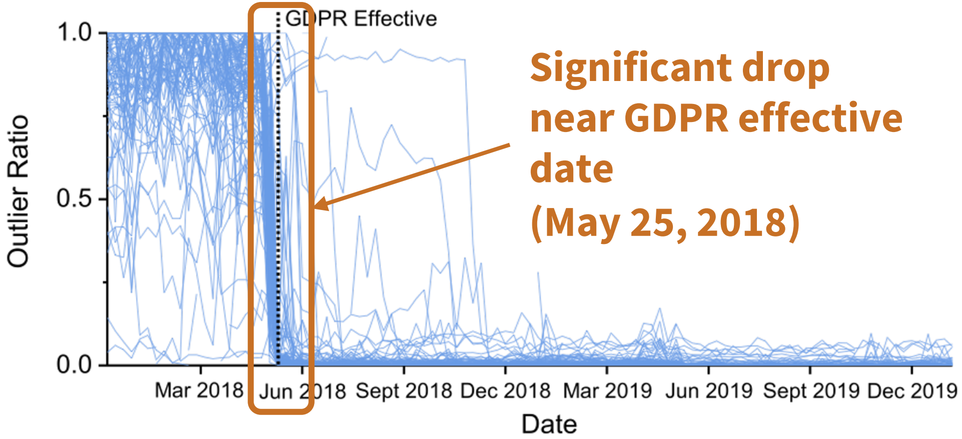
Over 80% fully-compliant providers completed in time

100 / 124 completed before May 25, 2018

Prominent efforts were taken *after* TempSpec (May 17)

Providers lack specific guidelines, thus chose to wait

Only *1 week* left for providers to take actions



Significant drop near GDPR effective date (May 25, 2018)

Measures of WHOIS Data Redaction

Contact masking measures

TempSpec: Use redacted value / empty value / privacy protection services

Category	# Provider	Example provider and values
Redacted value	**58**	ID-69 Tucows Domains Inc. (*"Redacted for privacy"*)
		ID-2 Network Solutions, LLC (*"statutory masking enabled"*)
		ID-625 Name.com, Inc. (*"non-public data"*)
		ID-1505 Gransy, s.r.o. (*"not disclosed"*)
Empty value	**63**	ID-146 GoDaddy.com, LLC; Public Internet Registry (PIR)
Privacy protection	**13**	ID-1456 NetArt Registrar Sp. z o.o. (*whoisdataprotection.com*)

Measures of WHOIS Data Redaction

Email anonymization measures

TempSpec: Use web form / anonymized email that _facilitate communication_

Over 25% fully-compliant registrars _do not_ offer such channel

Facilitates Communication	# Registrar	Interface	Example
Yes	**42 (72%)**	Web form	(_https://www.godaddy.com/whois/results.aspx_)
		Email	(_f***************7@proxyregistrant.email_)
No	**21 (28%)**	Web	(_https://tieredaccess.com_)
		Email	(_abuse@web.com_)

Scope of WHOIS Data Redaction

TempSpec lets providers decide what data to protect

- Apply to EEA domains only / Apply in a global basis

Scope of WHOIS Data Redaction

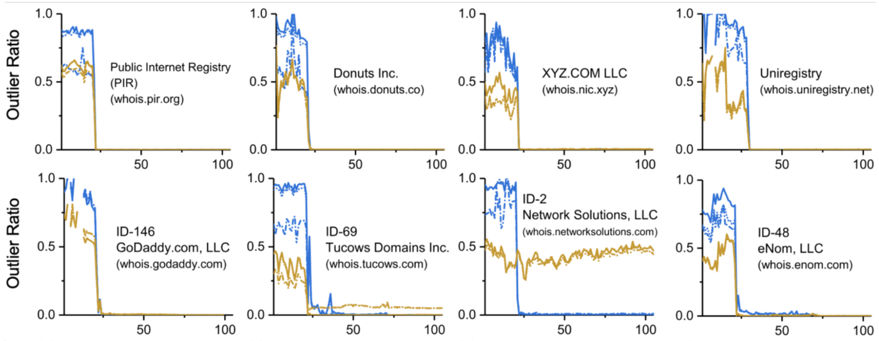
TempSpec lets providers decide what data to protect

Apply to EEA domains only / Apply in a global basis

Most providers sanitize *all* WHOIS data → Bad news for researchers

At least 60% fully-compliant providers apply globally

Causing a <u>*global, escalated loss*</u> of WHOIS



Comparison of outlier ratio of EEA and non-EEA records

Scope of WHOIS Data Redaction

TempSpec lets providers decide what data to protect

Apply to EEA domains only / Apply in a global basis

Most providers sanitize *all* WHOIS data → Bad news for researchers

At least 60% fully-compliant providers apply globally

Causing a *global, escalated loss* of WHOIS

Reasons?

1 week time is short for complete plans

Hard to determine what data is under scope

Saves work to comply with future policies (e.g., CCPA)

Part II:

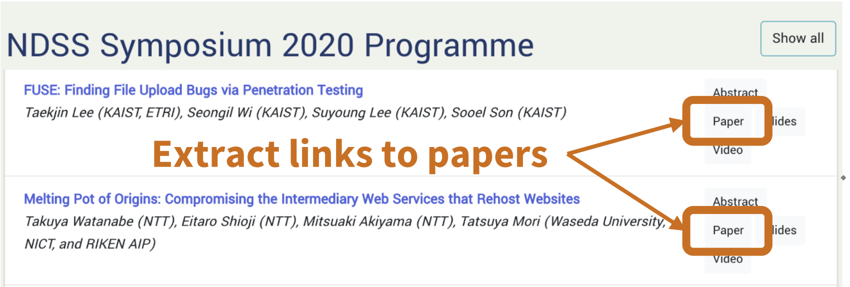
Security Impact of WHOIS Data Loss

WHOIS in Security Literature

Security papers published in 15 years of 5 conferences

NDSS, USENIX Security, IEEE S&P, ACM CCS, ACM IMC (2005 ~ 2020)

Download all via custom crawler



NDSS Symposium 2020 Programme

FUSE: Finding File Upload Bugs via Penetration Testing

Taekjin Lee (KAIST, ETRI), Seongil Wi (KAIST), Suyoung Lee (KAIST), Sooel Son (KAIST)

Melting Pot of Origins: Compromising the Intermediary Web Services that Rehost Websites

Takuya Watanabe (NTT), Eitaro Shioji (NTT), Mitsuaki Akiyama (NTT), Tatsuya Mori (Waseda University, NICT, and RIKEN AIP)

Extract links to papers

https://www.ndss-symposium.org/ndss-program/2020-program/

Paper database (4,304 paper PDFs)

Search keywords
(e.g., WHOIS, Domain)

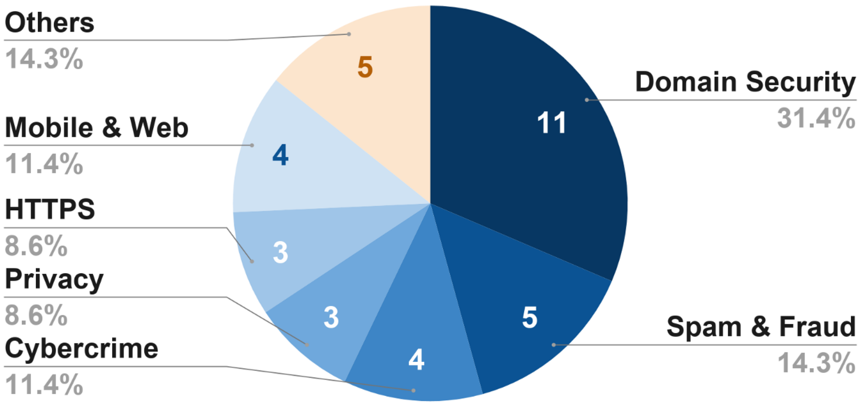
Manual confirmation

51 papers using WHOIS data

WHOIS in Security Literature

69% works that use WHOIS rely on redacted data

31 papers covering a wide range of security topics



Classified by security topics

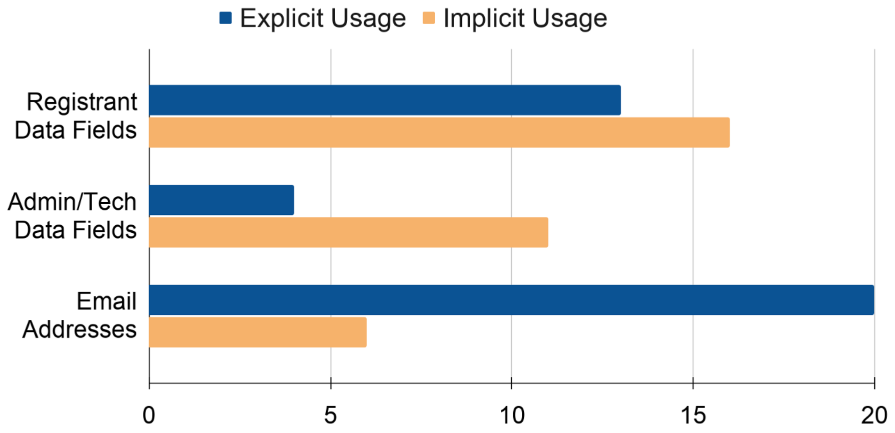
WHOIS Usage	Paper examples
Infer domain ownership / measurement purposes	Halvorson15, Vissers15, Chen16, Liu17
Features for detection	Sivakorn19, Le Pochat20
Vulnerability notification	Stock16, Stock18, Roth20
Result validation	Paxson13, Van Ede20, Delignat-Lavaud14,

WHOIS in Security Literature

69% works that use WHOIS rely on redacted data

31 papers covering a wide range of security topics

Registrant contact and *email addresses* are frequently used



Classified by WHOIS fields

Registrant contact: 29 papers (83%)

Admin/Tech contact: 15 papers (43%)

Email addresses: 26 papers (74%)

WHOIS in Security Literature

69% works that use WHOIS rely on redacted data

31 papers covering a wide range of security topics

Registrant contact and *email addresses* are frequently used

Other works not affected by WHOIS redaction

Use WHOIS fields that are not personal data

Creation date, Registrar info, Nameserver IP…

WHOIS in Security Literature

Hurdles for security researchers to access WHOIS

Over 70% WHOIS requests from security researchers are rejected*

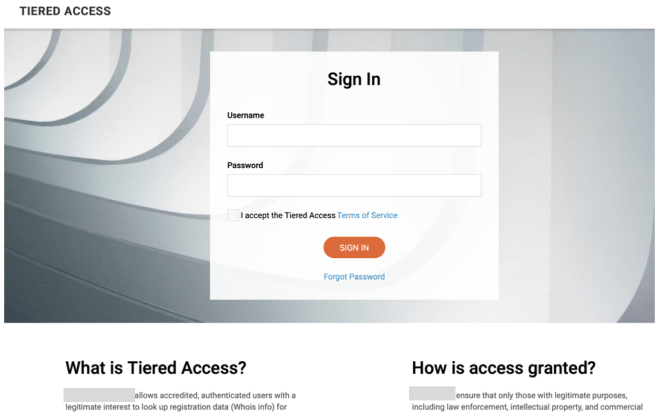
Current tiered systems lack instructions

Remediation: a better format of tiered access / data redaction

Use RDAP protocol to control access

Use Fuzzy hashing to replace fixed values

Review and adjust current security systems



TIERED ACCESS

Sign In

Username

Password

□ I accept the Tiered Access Terms of Service

SIGN IN

Forgot Password

What is Tiered Access?

[redacted] allows accredited, authenticated users with a legitimate interest to look up registration data (Whois info) for

How is access granted?

[redacted] ensure that only those with legitimate purposes, including law enforcement, intellectual property, and commercial

(Tiered access system of a registrar)

* https://docs.apwg.org/reports/ICANN_GDPR_WHOIS_Users_Survey_20181018.pdf

Part III:

Discussion & Summary

Discussion

GDPR’s impact on WHOIS is substantial

Most WHOIS providers *actively* and *extensively* redact personal data

A number of security works are affected due to WHOIS loss

Lessons learnt: Enforcing privacy policies is still a complex task

TempSpec leaves flexibility for providers, but not enough time

Checking tools are helpful to identify implementation flaws

The task requires more efficient collaboration across communities

Recommendations

Recommendations to multiple stakeholders

Party	Recommendation
Tech and legal authorities	Allow more lead time for more efficient discussions
Internet Supervisors (e.g. ICANN)	Develop more specific guidelines to avoid confusion
WHOIS providers	Review data protection implementations
Security researchers	Review and adjust security systems that rely on WHOIS

Search Engine for Security Papers

Search published security papers by keywords

Location: https://secpaper.cn/about

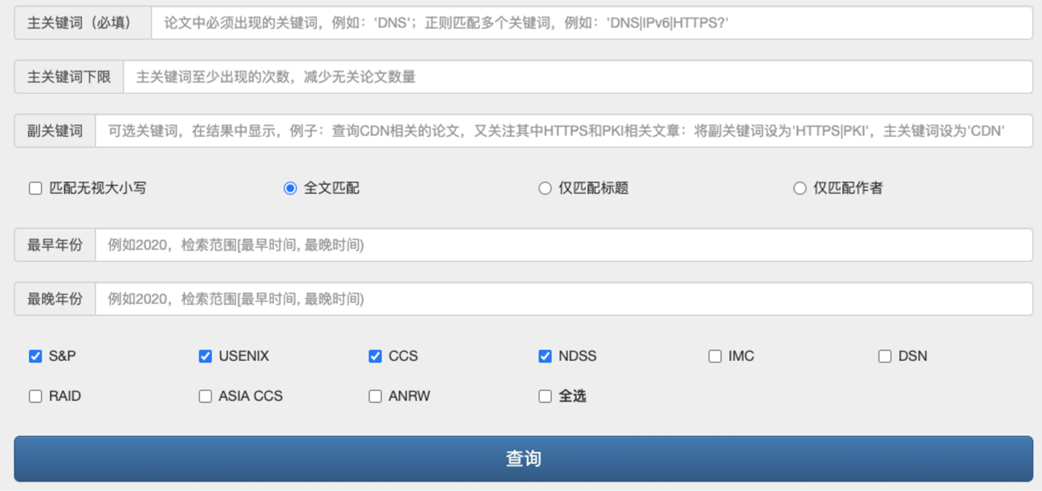
Conferences: IEEE S&P, USENIX Security, CCS, NDSS, IMC, DSN, RAID…

Trials and suggestions are welcome!

Credited to:

Fenglu Zhang @ Tsinghua

zfl20@mails.tsinghua.edu.cn



主关键词（必填） 论文中必须出现的关键词，例如：'DNS'；正则匹配多个关键词，例如：'DNS|IPv6|HTTPS?'

主关键词下限 主关键词至少出现的次数，减少无关论文数量

副关键词 可选关键词，在结果中显示，例子：查询CDN相关的论文，又关注其中HTTPS和PKI相关文章：将副关键词设为'HTTPS|PKI'，主关键词设为'CDN'

☐ 匹配无视大小写 ● 全文匹配 ○ 仅匹配标题 ○ 仅匹配作者

最早年份 例如2020，检索范围[最早时间, 最晚时间)

最晚年份 例如2020，检索范围[最早时间, 最晚时间)

☑ S&P ☑ USENIX ☑ CCS ☑ NDSS ☐ IMC ☐ DSN

☐ RAID ☐ ASIA CCS ☐ ANRW ☐ 全选

Summary

GDPR's impact is profound on WHOIS

Large WHOIS providers *actively* and *extensively* redact WHOIS data

Implementation flaws need to be fixed

The *excessive data protection scope* causes global WHOIS loss

A wide range of security works need review or adjustment

Redacted WHOIS data is widely used by security literature

Lessons learnt

Multiple stakeholders need more efficient collaboration

Release compliance checking tool

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