Protecting Privacy in Software Logs: What Should be Anonymized?

Roozbeh Aghili, Heng Li, Foutse Khomh

MOOSE

POLYTECHNIQUE
MONTRÉAL

Before going into this study, a progress report:

Reviewers' opinions!

REVIEWER 1

----------- Originality - ----------
SCORE: 3 (reasonably original)
----------- Soundness -----------
SCORE: 3 (some reservations)
--- -------- Relevance -----------
SCORE: 3 (some relevance)
----------- Reproducibility -----------
SCORE: 3 (some information not clear)
----------- Presentation -----------
SCORE: 3 (fair presentation)
----------- Overall evaluation -----------
SCORE: -1 (weak reject)

REVIEWER 2

----------- Originality -----------
SCORE: 1 (not at all original)
----------- Soundness -----------
SCORE: 3 (some reservations)
----------- Relevance -------- ---
SCORE: 2 (very little relevance)
----------- Reproducibility -----------
SCORE: 3 (some information not clear)
------- ---- Presentation -----------
SCORE: 4 (good presentation)
----------- Overall evaluation -----------
SCORE: - 2 (reject)

REVIEWER 3

----------- Originality -----------
SCORE: 2 (somewhat original)
----------- Soundness -----------
SCORE: 3 (some reservations)
----------- Relevance -----------
SCORE: 4 (relevant)
----------- Reproducibility -----------
SCORE: 3 (some information not clear)
-- --------- Presentation -----------
SCORE: 4 (good presentation)
---------- Overall evaluation --------- --
SCORE: -1 (weak reject)

Previous version

Characterizing the Workload Patterns of Web Applications

RQ1. What are the **existing workload patterns** in web application traces?

RQ2. How are different workload patterns **distributed** in web application traces?

Current version

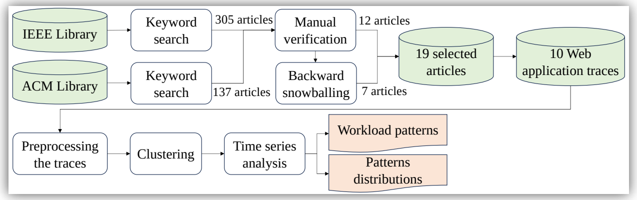
Understanding Web Application Workloads and Their Applications

RQ1. What are the **applications** of web application workloads in **existing works**?

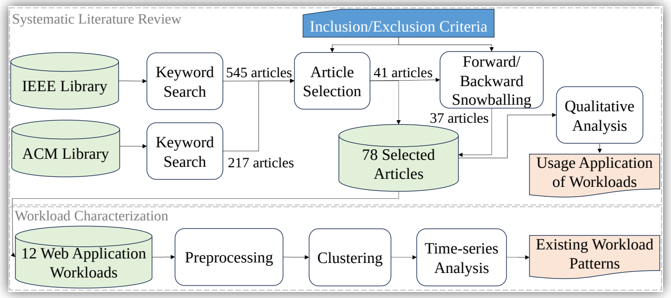
RQ2. What are the **existing patterns** in web application workloads?

Previous version

Overview of our study

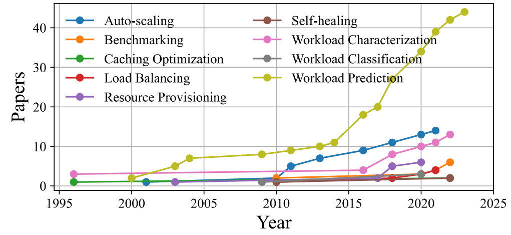


Current version

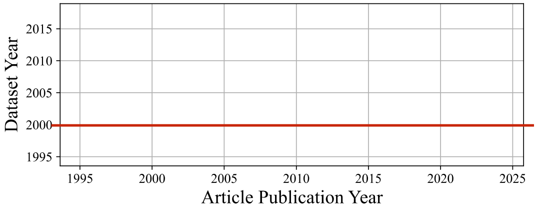


Current version

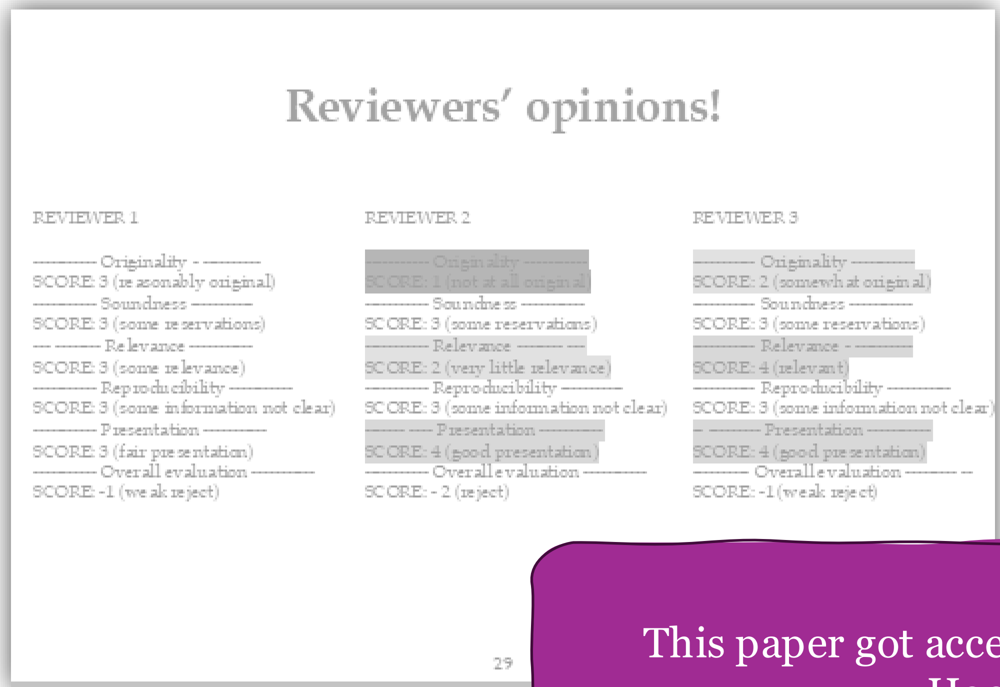
Because of our systematic literature review



Cumulative number of papers per objective over the years



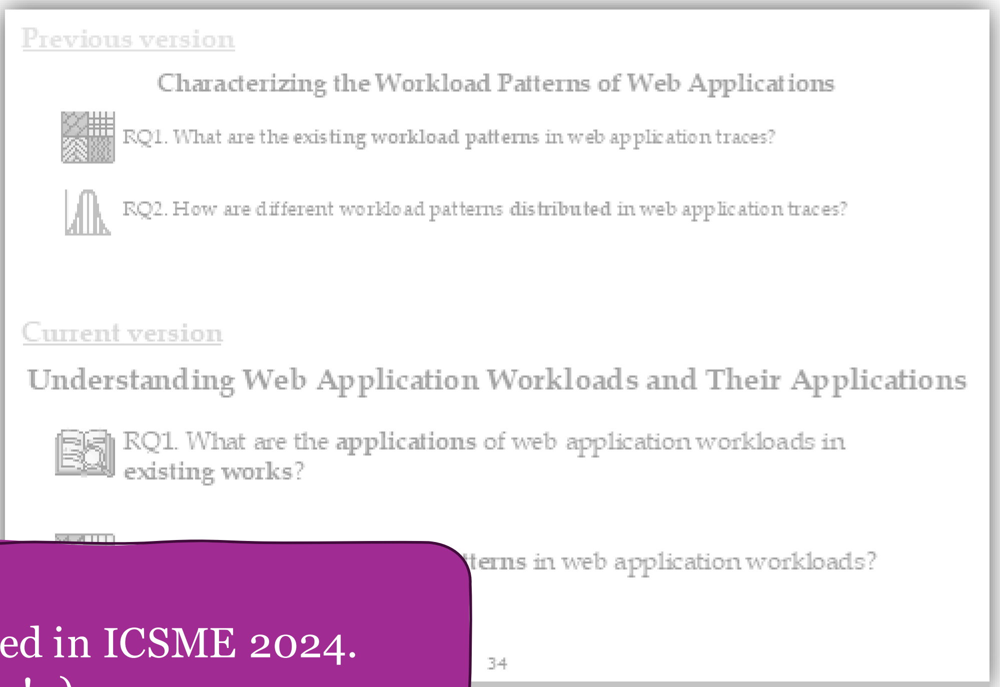
Comparison of literature publication years and corresponding workload dataset years



Reviewers' opinions!

REVIEWER 1	REVIEWER 2	REVIEWER 3
---------- Originality - ----------	---------- Originality ----------	---------- Originality ----------
SCORE: 3 (reasonably original)	SCORE: 1 (not at all original)	SCORE: 2 (somewhat original)
---------- Soundness ----------	---------- Soundness ----------	---------- Soundness ----------
SCORE: 3 (some reservations)	SCORE: 3 (some reservations)	SCORE: 3 (some reservations)
--- -------- Relevance ----------	---------- Relevance -------- ---	---------- Relevance ----------
SCORE: 3 (some relevance)	SCORE: 2 (very little relevance)	SCORE: 4 (relevant)
---------- Reproducibility ----------	---------- Reproducibility ----------	---------- Reproducibility ----------
SCORE: 3 (some information not clear)	SCORE: 3 (some information not clear)	SCORE: 3 (some information not clear)
---------- Presentation ----------	------ ---- Presentation ----------	-- -------- Presentation ----------
SCORE: 3 (fair presentation)	SCORE: 4 (good presentation)	SCORE: 4 (good presentation)
---------- Overall evaluation ----------	---------- Overall evaluation ----------	---------- Overall evaluation ---------- --
SCORE: -1 (weak reject)	SCORE: - 2 (reject)	SCORE: -1 (weak reject)

29



Previous version

Characterizing the Workload Patterns of Web Applications

RQ1. What are the existing workload patterns in web application traces?

RQ2. How are different workload patterns distributed in web application traces?

Current version

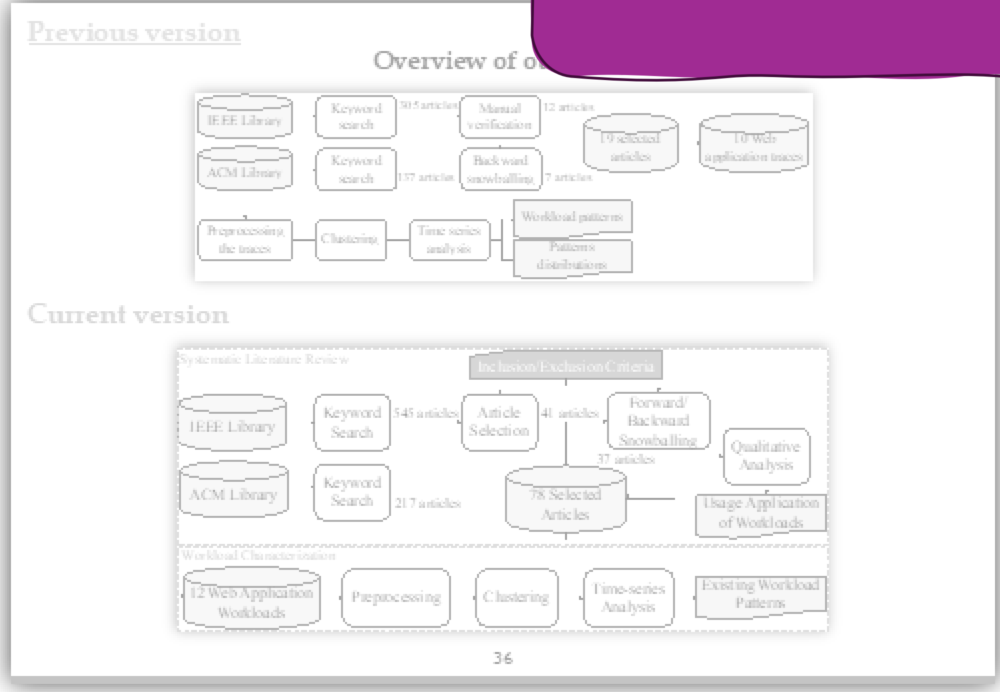
Understanding Web Application Workloads and Their Applications

RQ1. What are the applications of web application workloads in existing works?

…terns in web application workloads?

34

This paper got accepted in ICSME 2024.
Hooray! ;)

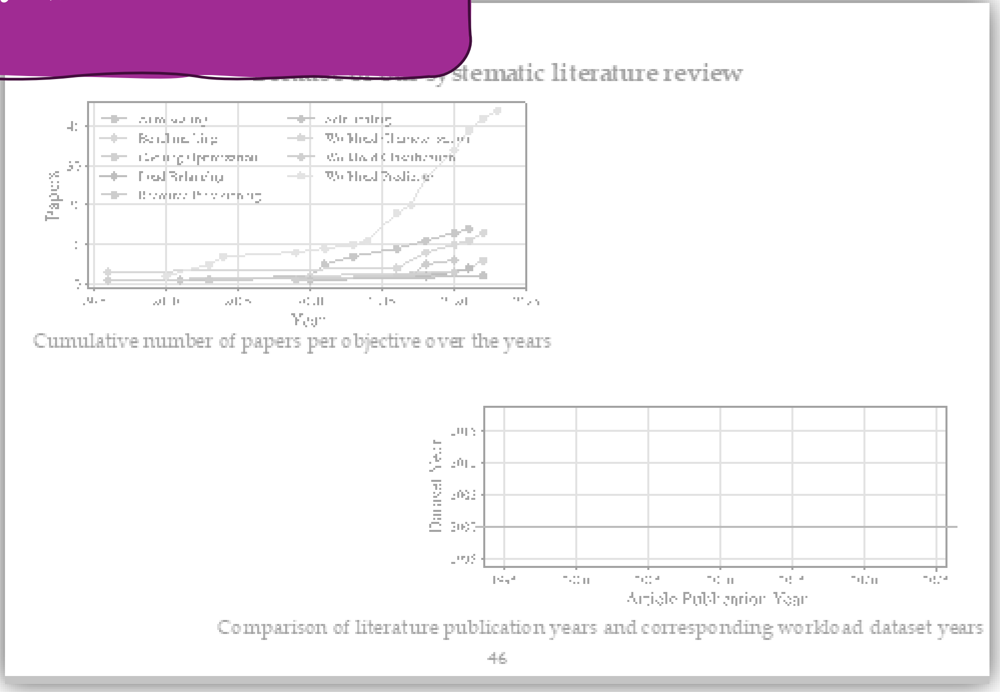


Previous version

Overview of o…

Current version

36



…stematic literature review

Cumulative number of papers per objective over the years

Comparison of literature publication years and corresponding workload dataset years

46

Before going into this study, a progress report:

1. **Aghili, R.**, Qin, Q., Li, H., & Khomh, F. (2024). Understanding Web Application Workloads and Their Applications: Systematic Literature Review and Characterization. *International Conference on Software Maintenance and Evolution (ICSME)* (accepted).

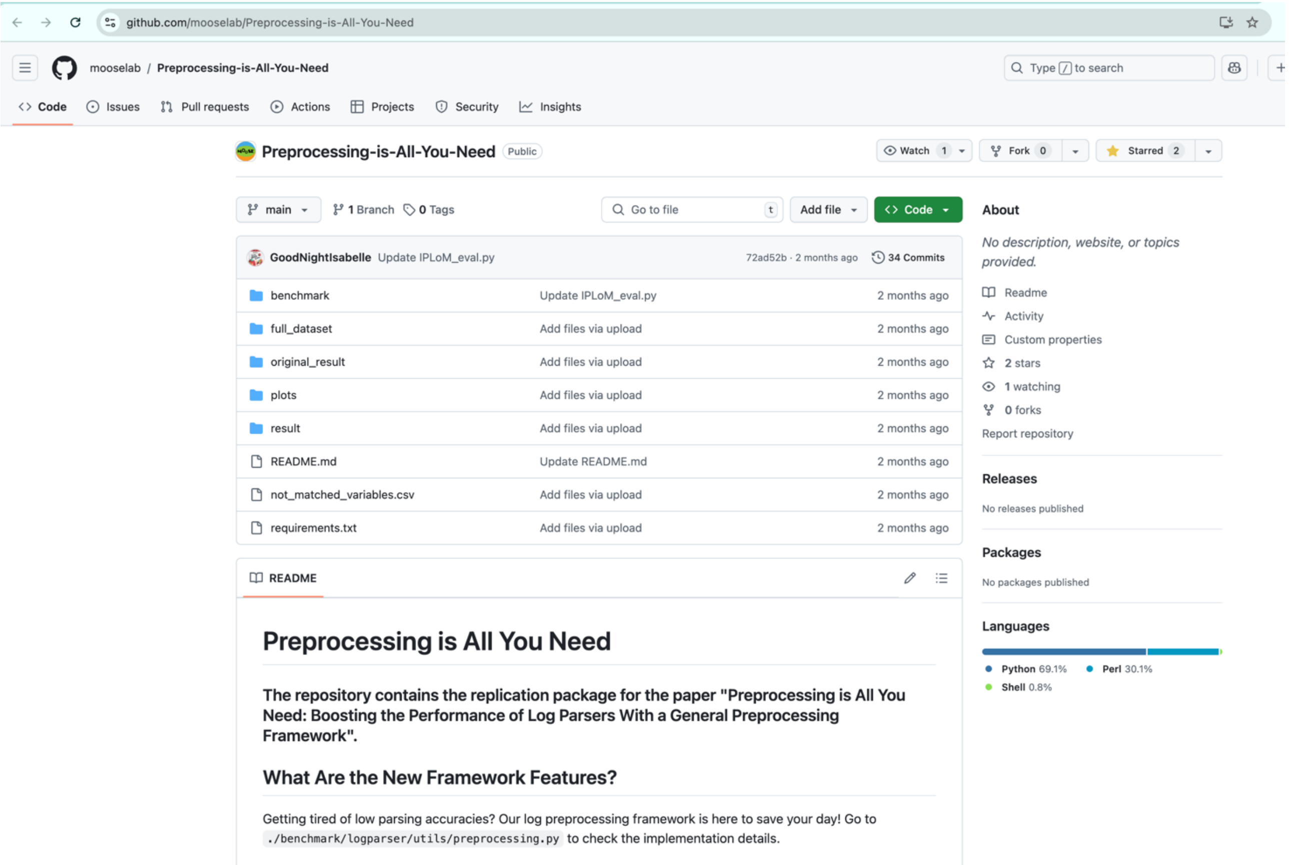
Dear Roozbeh,

Thank you for submitting to SANER 2025!

We are delighted to inform you that your submission

111 – Preprocessing is All You Need: Boosting the Performance of Log Parsers With a General Preprocessing Framework
Qiaolin Qin, Roozbeh Aghili, Heng Li, Ettore Merlo

has been accepted for inclusion in the SANER 2025 technical program . Congratulations!



GoodNightIsabelle Update IPLoM_eval.py 72ad52b · 2 months ago 34 Commits

benchmark	Update IPLoM_eval.py	2 months ago
full_dataset	Add files via upload	2 months ago
original_result	Add files via upload	2 months ago
plots	Add files via upload	2 months ago
result	Add files via upload	2 months ago
README.md	Update README.md	2 months ago
not_matched_variables.csv	Add files via upload	2 months ago
requirements.txt	Add files via upload	2 months ago

No description, website, or topics provided.

No releases published

No packages published

Preprocessing is All You Need

The repository contains the replication package for the paper "Preprocessing is All You Need: Boosting the Performance of Log Parsers With a General Preprocessing Framework".

What Are the New Framework Features?

Getting tired of low parsing accuracies? Our log preprocessing framework is here to save your day! Go to `./benchmark/logparser/utils/preprocessing.py` to check the implementation details.

Before going into this study, a progress report:

1. **Aghili, R**., Qin, Q., Li, H., & Khomh, F. (2024). Understanding Web Application Workloads and Their Applications: Systematic Literature Review and Characterization. *International Conference on Software Maintenance and Evolution (ICSME)* (accepted).
2. Qin, Q., **Aghili, R**., Li, H., & Merlo, E. (2025). Preprocessing is All You Need: Boosting the Performance of Log Parsers With a General Preprocessing Framework. International Conference on Software Analysis, Evolution and Reengineering (SANER) (accepted).

Protecting Privacy in Software Logs: What Should be Anonymized?

Roozbeh Aghili, Heng Li, Foutse Khomh

Submitted to FSE 2025.

Before going into this study, a progress report:

1. **Aghili, R.**, Qin, Q., Li, H., & Khomh, F. (2024). Understanding Web Application Workloads and Their Applications: Systematic Literature Review and Characterization. *International Conference on Software Maintenance and Evolution (ICSME)* (accepted).
2. Qin, Q., **Aghili, R.**, Li, H., & Merlo, E. (2025). Preprocessing is All You Need: Boosting the Performance of Log Parsers With a General Preprocessing Framework. International Conference on Software Analysis, Evolution and Reengineering (SANER) (accepted).
3. **Aghili, R.**, Li, H., & Khomh, F. (2025). Protecting Privacy in Software Logs: What Should be Anonymized? International Conference on the Foundations of Software Engineering (FSE) (submitted).

yahoo!

yahoo!

2013-2014

yahoo!

2013-2014

Over 3 billion accounts

yahoo!

2013-2014

- Names
- Email addresses
- Phone numbers
- Birth dates
- Passwords
- Calendars
- Security questions

Over 3 billion accounts

yahoo!

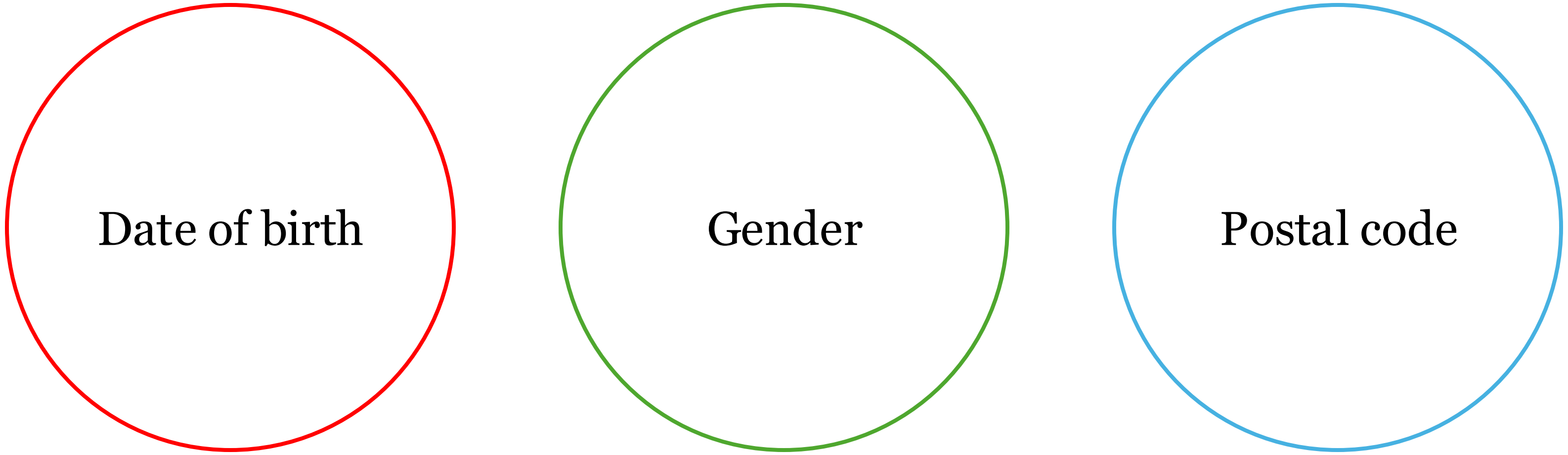
2013-2014

- Names
- Email addresses
- Phone numbers
- Birth dates
- Passwords
- Calendars
- Security questions

Over 3 billion accounts

Personally Identifiable Information (PII)

- Name
- Email address
- Phone number
- Security number
- Driver’s license number



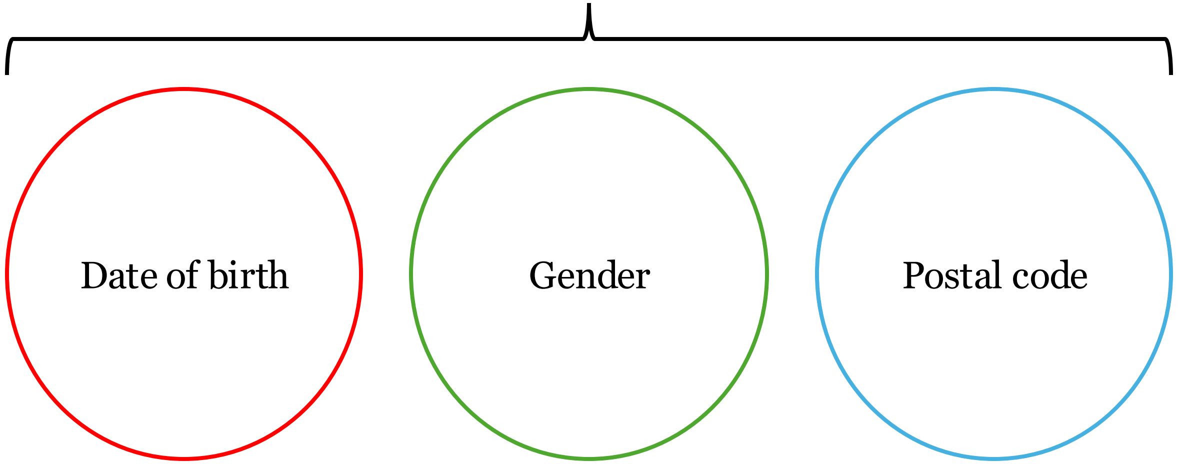
Date of birth

Gender

Postal code

[1] Sweeney, L. (2002), "K-anonymity: A model for protecting privacy”

Quasi-Identifiers



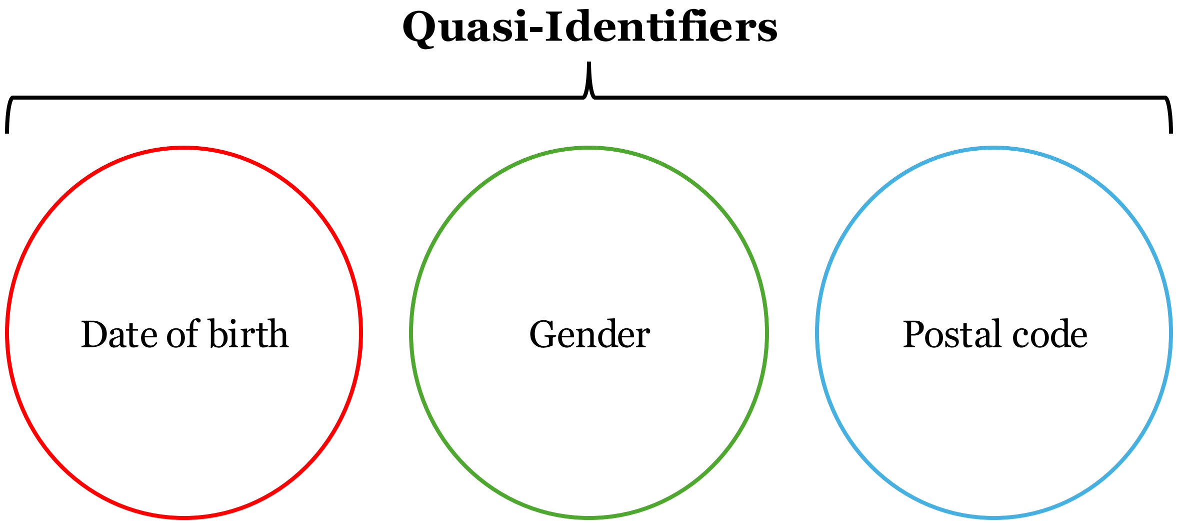
Date of birth

Gender

Postal code

87% of US population

[1] Sweeney, L. (2002), "K-anonymity:
A model for protecting privacy”



Quasi-Identifiers

Date of birth

Gender

Postal code

87% of US population

William Weld
68th Governor of
Massachusetts

[1] Sweeney, L. (2002), "K-anonymity: A model for protecting privacy”

Personally Identifiable Information (PII)

- Name
- Email address
- Phone number
- Security number
- Driver’s license number

Quasi-Identifiers

Date of birth

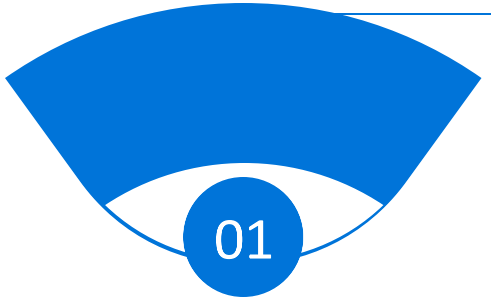
Gender

Postal code

But how about SOFTWARE LOGS?

What Should be Anonymized?

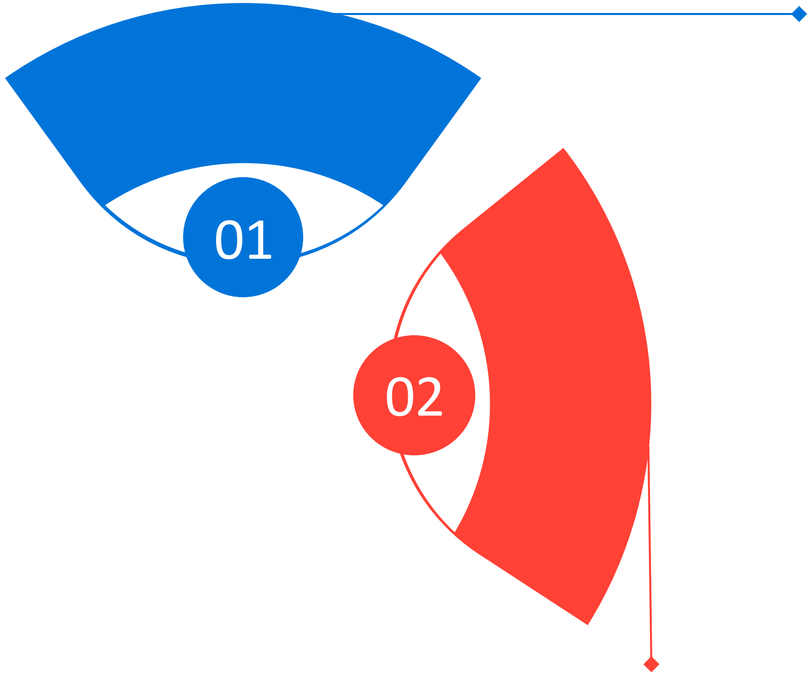
What Should be Anonymized?



Software Logs

25 log datasets:
1. Loghub datasets
2. Web app datasets

What Should be Anonymized?



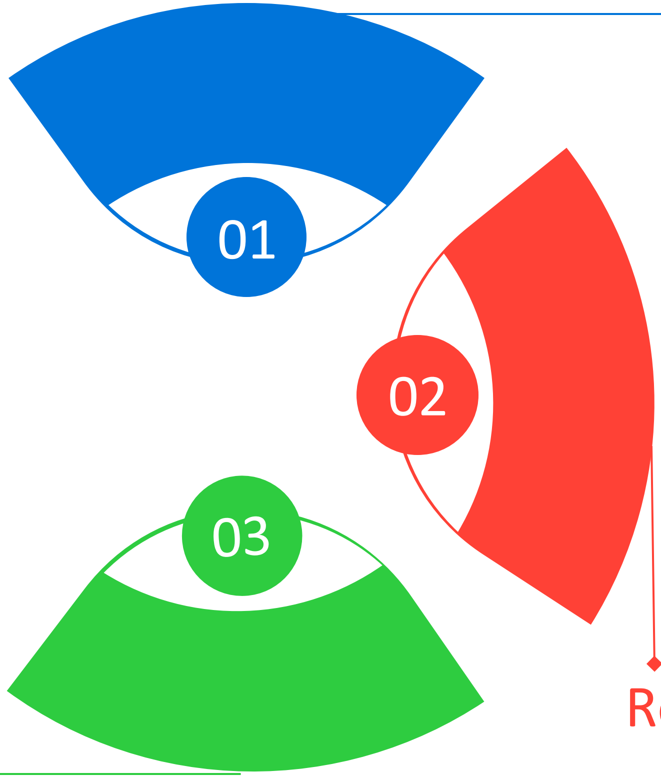
Software Logs

25 log datasets:
1. Loghub datasets
2. Web app datasets

Regulations

GDPR (EU)
HIPAA (US)
CCPA (US)
PIPEDA (CA)
ISO27001

What Should be Anonymized?



Software Logs

25 log datasets:
1. Loghub datasets
2. Web app datasets

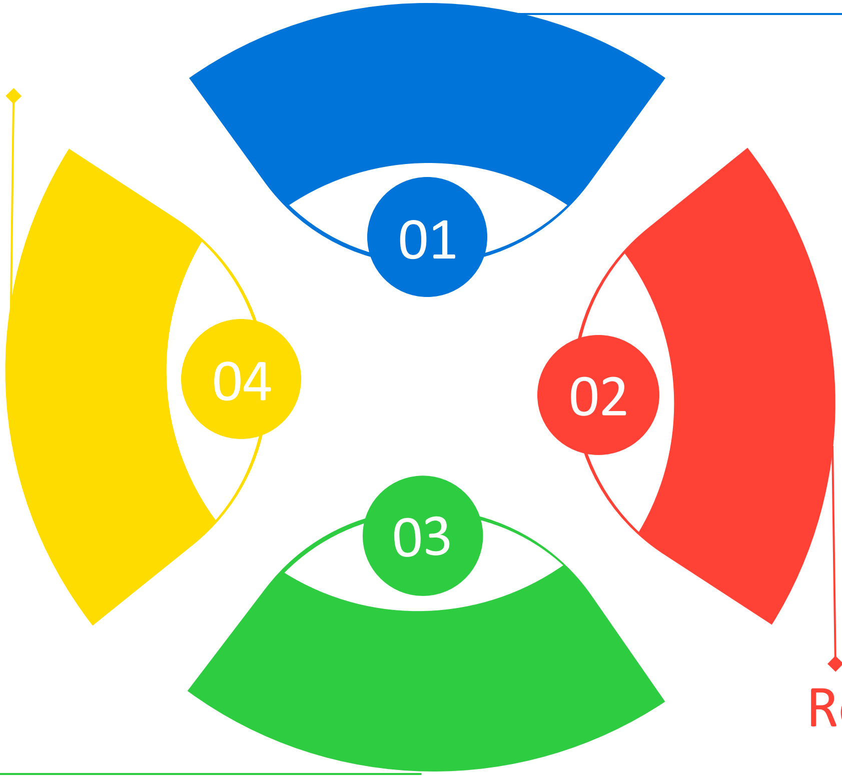
Regulations

GDPR (EU)
HIPAA (US)
CCPA (US)
PIPEDA (CA)
ISO27001

Articles & Tools

58 articles
1. Search on 2 libraries:
IEEE and ACM
2. Snowballing

What Should be Anonymized?



Survey

45 industry participants

Software Logs

25 log datasets:
1. Loghub datasets
2. Web app datasets

Articles & Tools

58 articles
1. Search on 2 libraries:
IEEE and ACM
2. Snowballing

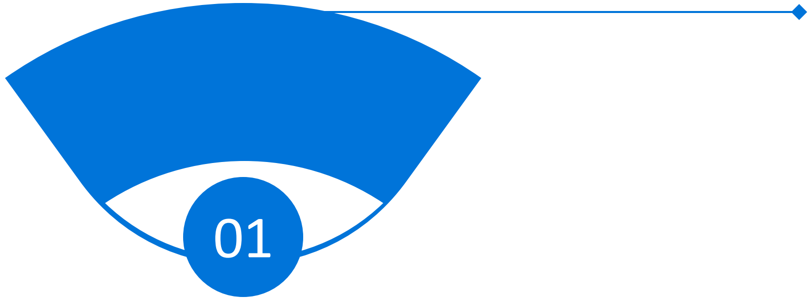
Regulations

GDPR (EU)
HIPAA (US)
CCPA (US)
PIPEDA (CA)
ISO27001

01

Software Logs

25 log datasets:
1. Loghub datasets
2. Web app datasets

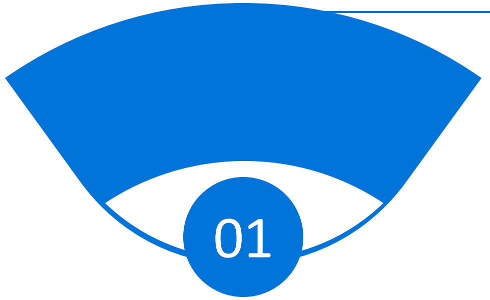


Software Logs

25 log datasets:
1. Loghub datasets
2. Web app datasets

Different types of log datasets:

1. Distributed systems
2. Super computers
3. Operating systems
4. Mobile systems
5. Server applications
6. Standalone software
7. Web applications



Software Logs

25 log datasets:
1. Loghub datasets
2. Web app datasets

For each dataset:

1. Sample 2000 lines
2. Parse using Drain

Different types of log datasets:

1. Distributed systems
2. Super computers
3. Operating systems
4. Mobile systems
5. Server applications
6. Standalone software
7. Web applications

01

Software Logs

25 log datasets:
1. Loghub datasets
2. Web app datasets

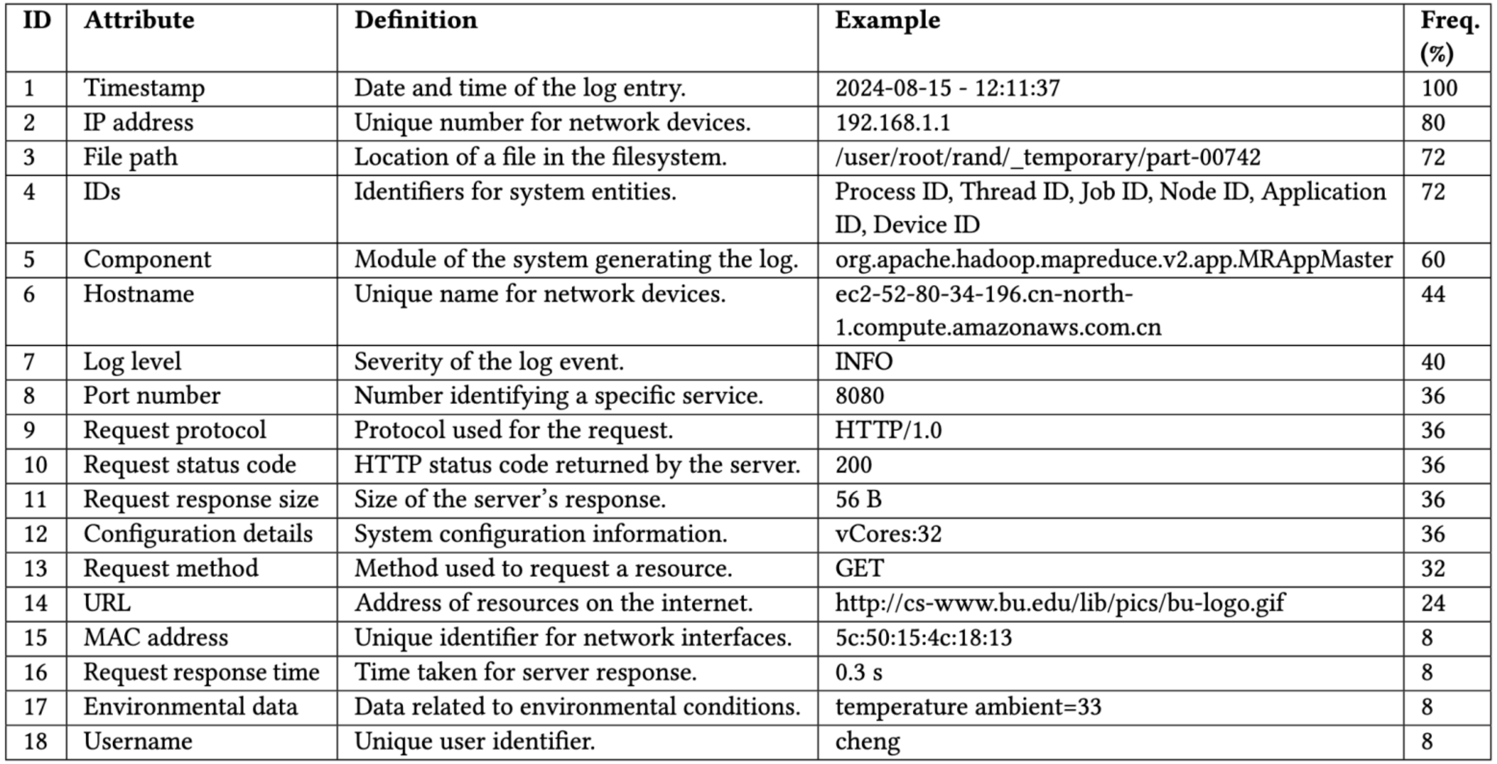
For each dataset:

1. Sample 2000 lines
2. Parse using Drain

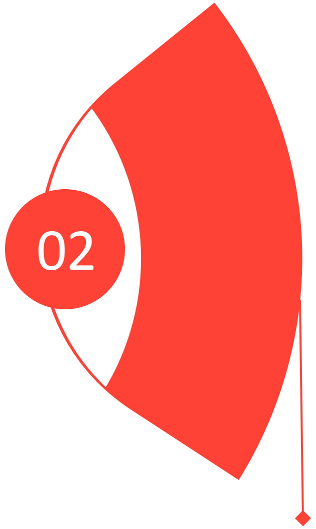
Different types of log datasets:

1. Distributed systems
2. Super computers
3. Operating systems
4. Mobile systems
5. Server applications
6. Standalone software
7. Web applications

Table 2. The most frequent log attributes and examples



ID	Attribute	Definition	Example	Freq. (%)
1	Timestamp	Date and time of the log entry.	2024-08-15 - 12:11:37	100
2	IP address	Unique number for network devices.	192.168.1.1	80
3	File path	Location of a file in the filesystem.	/user/root/rand/_temporary/part-00742	72
4	IDs	Identifiers for system entities.	Process ID, Thread ID, Job ID, Node ID, Application ID, Device ID	72
5	Component	Module of the system generating the log.	org.apache.hadoop.mapreduce.v2.app.MRAppMaster	60
6	Hostname	Unique name for network devices.	ec2-52-80-34-196.cn-north-1.compute.amazonaws.com.cn	44
7	Log level	Severity of the log event.	INFO	40
8	Port number	Number identifying a specific service.	8080	36
9	Request protocol	Protocol used for the request.	HTTP/1.0	36
10	Request status code	HTTP status code returned by the server.	200	36
11	Request response size	Size of the server's response.	56 B	36
12	Configuration details	System configuration information.	vCores:32	36
13	Request method	Method used to request a resource.	GET	32
14	URL	Address of resources on the internet.	http://cs-www.bu.edu/lib/pics/bu-logo.gif	24
15	MAC address	Unique identifier for network interfaces.	5c:50:15:4c:18:13	8
16	Request response time	Time taken for server response.	0.3 s	8
17	Environmental data	Data related to environmental conditions.	temperature ambient=33	8
18	Username	Unique user identifier.	cheng	8



02

Regulations

GDPR (EU)
HIPAA (US)
CCPA (US)
PIPEDA (CA)
ISO27001

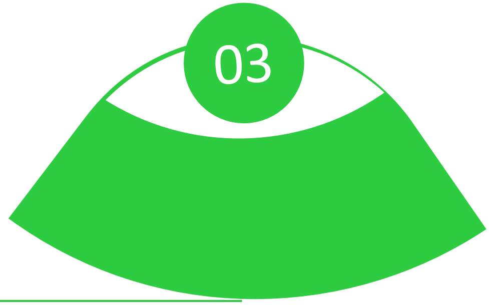
02

Regulations

GDPR (EU)
HIPAA (US)
CCPA (US)
PIPEDA (CA)
ISO27001

Summary

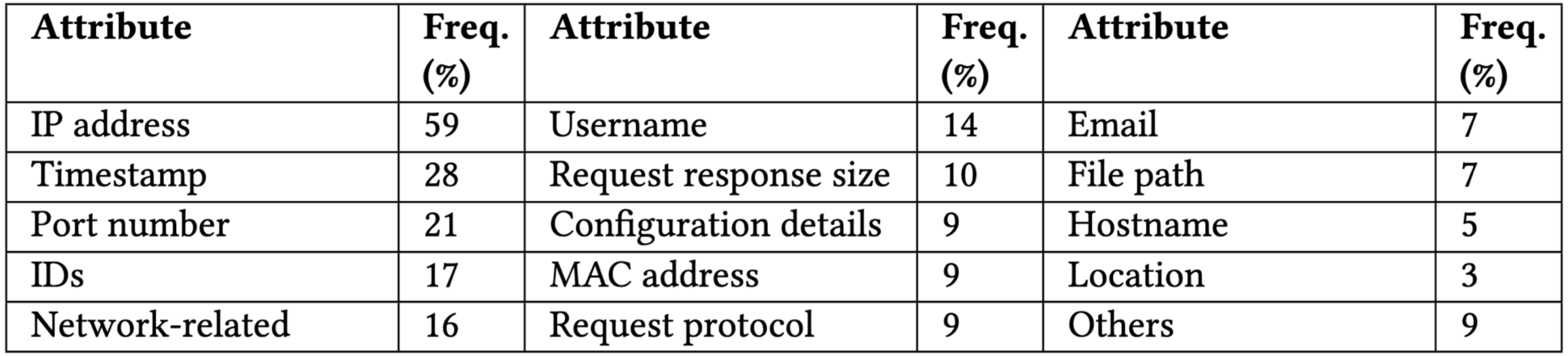
While numerous data privacy regulations exist, none specifically address software logs. Therefore, it is essential to extract relevant information from these regulations that could be applicable to logs. Some regulations explicitly define personal data and specify attributes that need protection. For instance, GDPR and HIPAA both classify IP addresses as sensitive data. In contrast, ISO 27001 does not define personal or sensitive data but instead offers a flexible framework for managing any data that an organization identifies as sensitive.



Articles & Tools

58 articles
1. Search on 2 libraries:
IEEE and ACM
2. Snowballing

Table 3. Usage of sensitive log attributes in reviewed articles



Attribute	Freq. (%)	Attribute	Freq. (%)	Attribute	Freq. (%)
IP address	59	Username	14	Email	7
Timestamp	28	Request response size	10	File path	7
Port number	21	Configuration details	9	Hostname	5
IDs	17	MAC address	9	Location	3
Network-related	16	Request protocol	9	Others	9

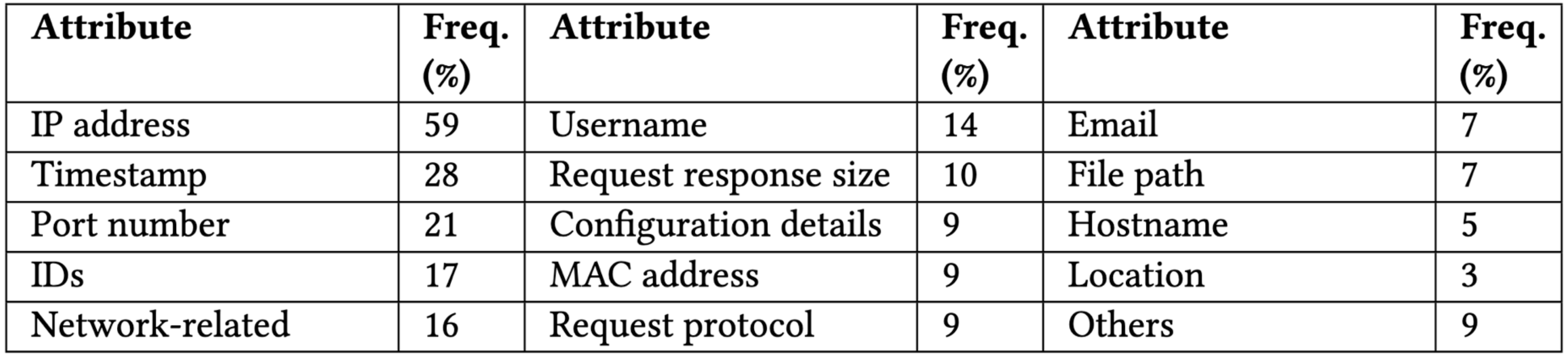
03

Articles & Tools

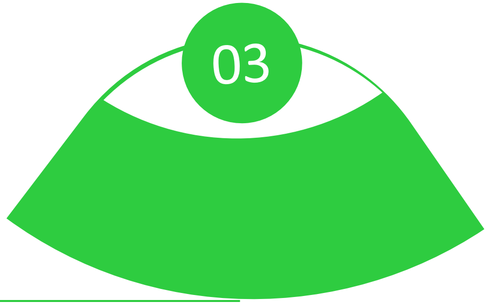
58 articles
1. Search on 2 libraries: IEEE and ACM
2. Snowballing

Many studies only focus on the privacy of IP addresses.
Many studies only focus on the network-related attributes.

Table 3. Usage of sensitive log attributes in reviewed articles



Attribute	Freq. (%)	Attribute	Freq. (%)	Attribute	Freq. (%)
IP address	59	Username	14	Email	7
Timestamp	28	Request response size	10	File path	7
Port number	21	Configuration details	9	Hostname	5
IDs	17	MAC address	9	Location	3
Network-related	16	Request protocol	9	Others	9

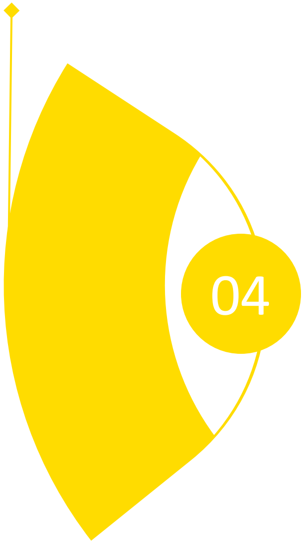


Articles & Tools

58 articles
1. Search on 2 libraries: IEEE and ACM
2. Snowballing

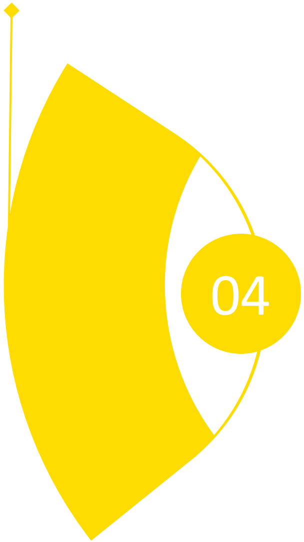
Survey

45 industry participants



Survey

45 industry participants

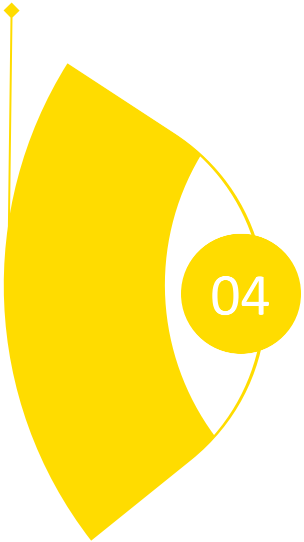


Different types of questions:

1. Multiple-choice
2. Likert-scale
3. Open-ended
4. Demographic

Survey

45 industry participants



Different types of questions:

1. Multiple-choice
2. Likert-scale
3. Open-ended
4. Demographic

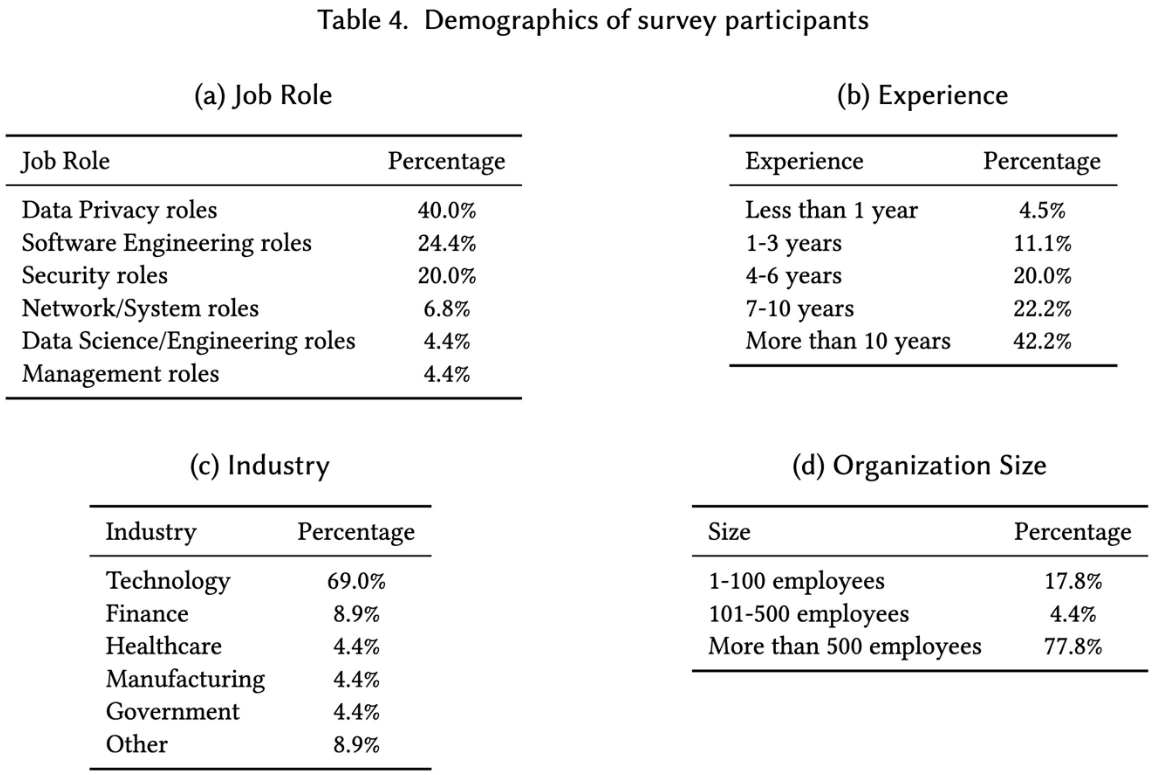


Table 4. Demographics of survey participants

(a) Job Role

Job Role	Percentage
Data Privacy roles	40.0%
Software Engineering roles	24.4%
Security roles	20.0%
Network/System roles	6.8%
Data Science/Engineering roles	4.4%
Management roles	4.4%

(b) Experience

Experience	Percentage
Less than 1 year	4.5%
1-3 years	11.1%
4-6 years	20.0%
7-10 years	22.2%
More than 10 years	42.2%

(c) Industry

Industry	Percentage
Technology	69.0%
Finance	8.9%
Healthcare	4.4%
Manufacturing	4.4%
Government	4.4%
Other	8.9%

(d) Organization Size

Size	Percentage
1-100 employees	17.8%
101-500 employees	4.4%
More than 500 employees	77.8%

Survey

45 industry participants

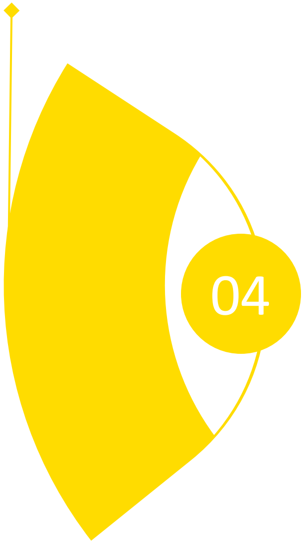
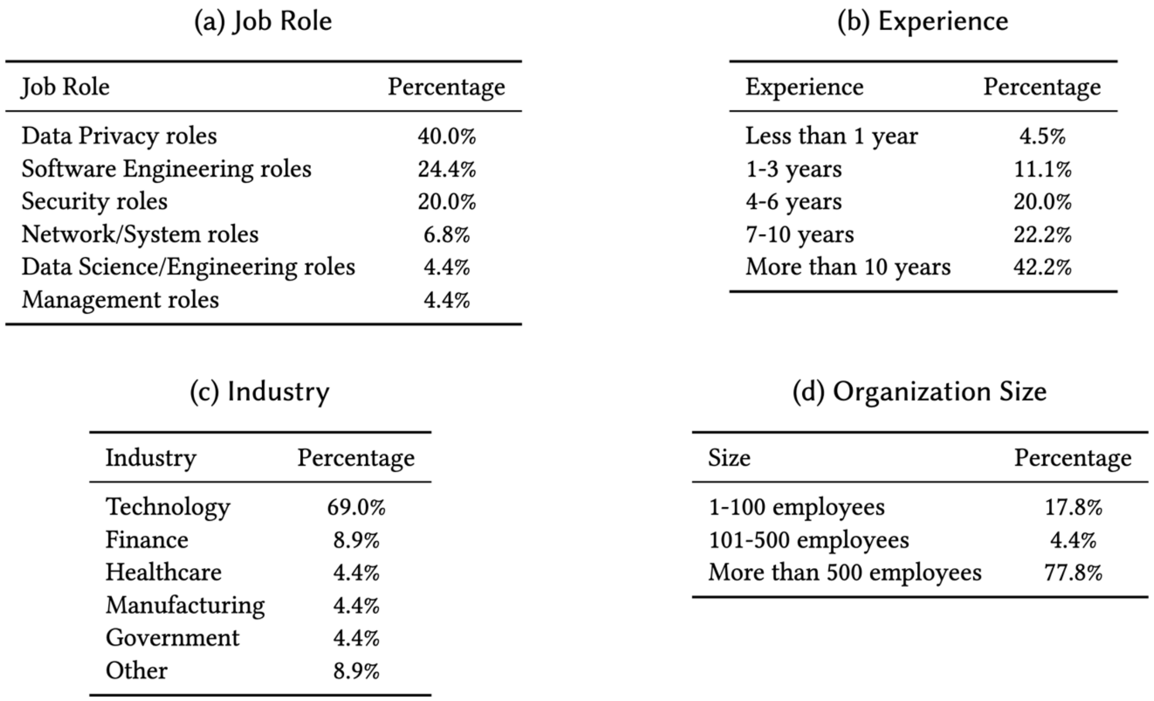


Table 4. Demographics of survey participants



(a) Job Role

Job Role	Percentage
Data Privacy roles	40.0%
Software Engineering roles	24.4%
Security roles	20.0%
Network/System roles	6.8%
Data Science/Engineering roles	4.4%
Management roles	4.4%

(b) Experience

Experience	Percentage
Less than 1 year	4.5%
1-3 years	11.1%
4-6 years	20.0%
7-10 years	22.2%
More than 10 years	42.2%

(c) Industry

Industry	Percentage
Technology	69.0%
Finance	8.9%
Healthcare	4.4%
Manufacturing	4.4%
Government	4.4%
Other	8.9%

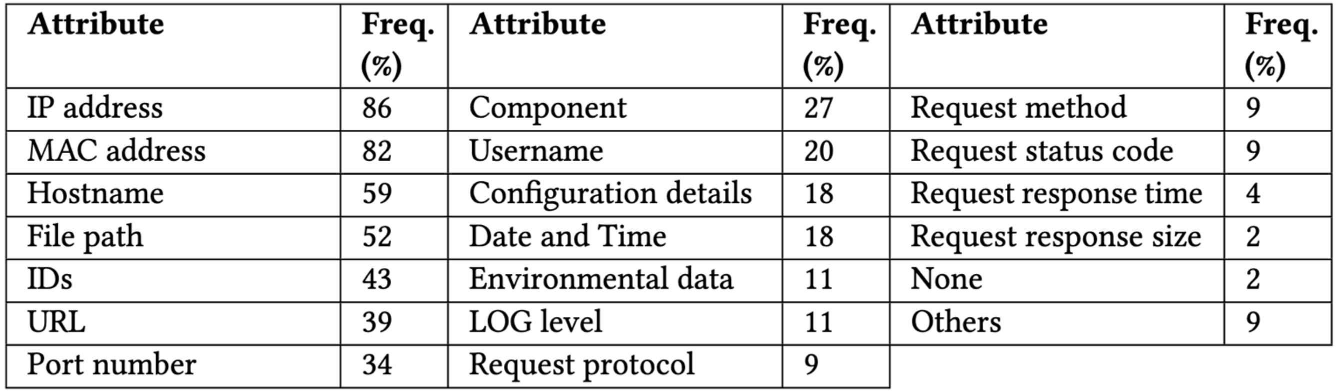
(d) Organization Size

Size	Percentage
1-100 employees	17.8%
101-500 employees	4.4%
More than 500 employees	77.8%

Different types of questions:

1. Multiple-choice
2. Likert-scale
3. Open-ended
4. Demographic

Table 5. The sensitive log attributes from industry perspective



Attribute	Freq. (%)	Attribute	Freq. (%)	Attribute	Freq. (%)
IP address	86	Component	27	Request method	9
MAC address	82	Username	20	Request status code	9
Hostname	59	Configuration details	18	Request response time	4
File path	52	Date and Time	18	Request response size	2
IDs	43	Environmental data	11	None	2
URL	39	LOG level	11	Others	9
Port number	34	Request protocol	9		

Let’s see some examples!

Chinese University of Hong Kong (CUHK),
Department of Computer Science and Engineering

```
[10.30 16:49:06] chrome.exe - proxy.cse.cuhk.edu.hk:5070 open through proxy proxy.cse.cuhk.edu.hk:5070 HTTPS
```

Using Amazon services, server cn-north-1 (China, Beijing)

```
Dec 10 07:55:55 LabSZ sshd[24331]: pam_unix(sshd:auth): authentication failure; logname= uid=0 euid=0 tty=ssh ruser=
rhost=ec2-52-80-34-196.cn-north-1.compute.amazonaws.com.cn
```

Configuration details

```
2015-10-18 18:01:53,713 INFO [main] org.apache.hadoop.mapreduce.v2.app.rm.RMContainerAllocator:
maxContainerCapability: <memory:8192, vCores:32>
```

Ok, whatever, what should be anonymized finally?

Ok, whatever, what should be anonymized finally?

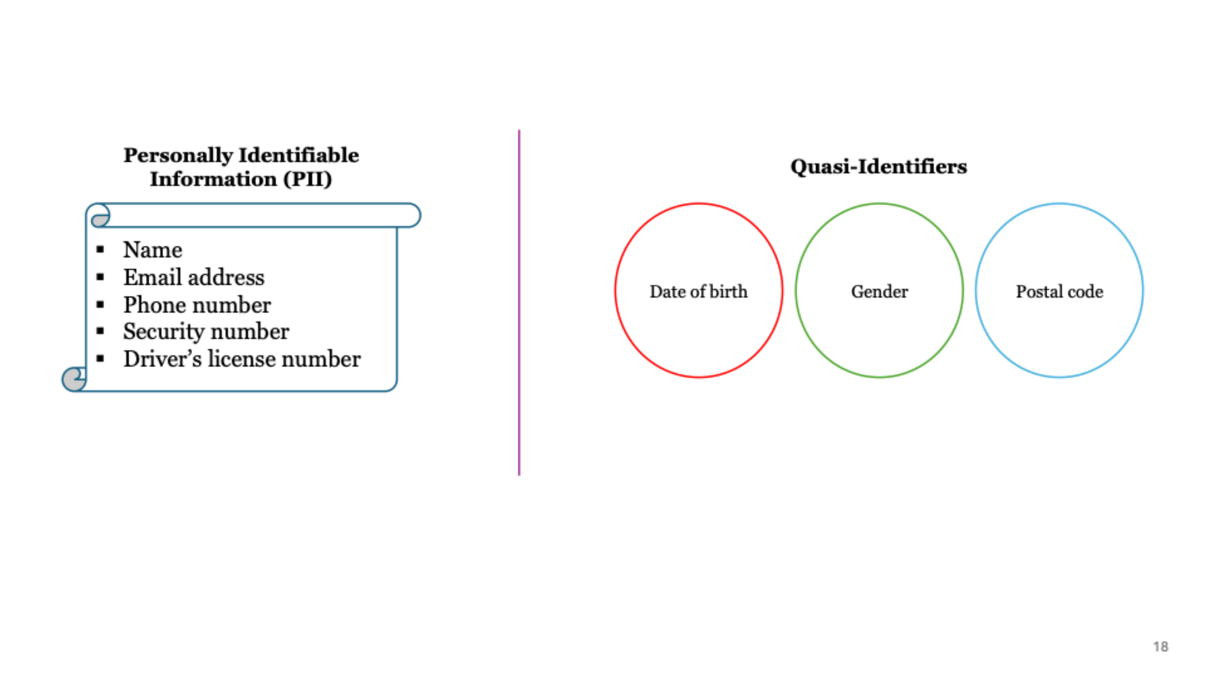
Based on our analyses of software log privacy from multiple perspectives, we consider these attributes as generally sensitive:

1. **IP addresses**
2. **MAC addresses**
3. **Hostnames**
4. **file paths**
5. **IDs**
6. **URLs**
7. **Usernames**
8. **Port numbers**
9. **Configuration details**

Research gaps and future directions?

Research gaps and future directions?

1. Broadening the focus on diverse log attributes.
2. Developing specialized anonymization tools for software logs.
3. Developing a privacy score for software logs.

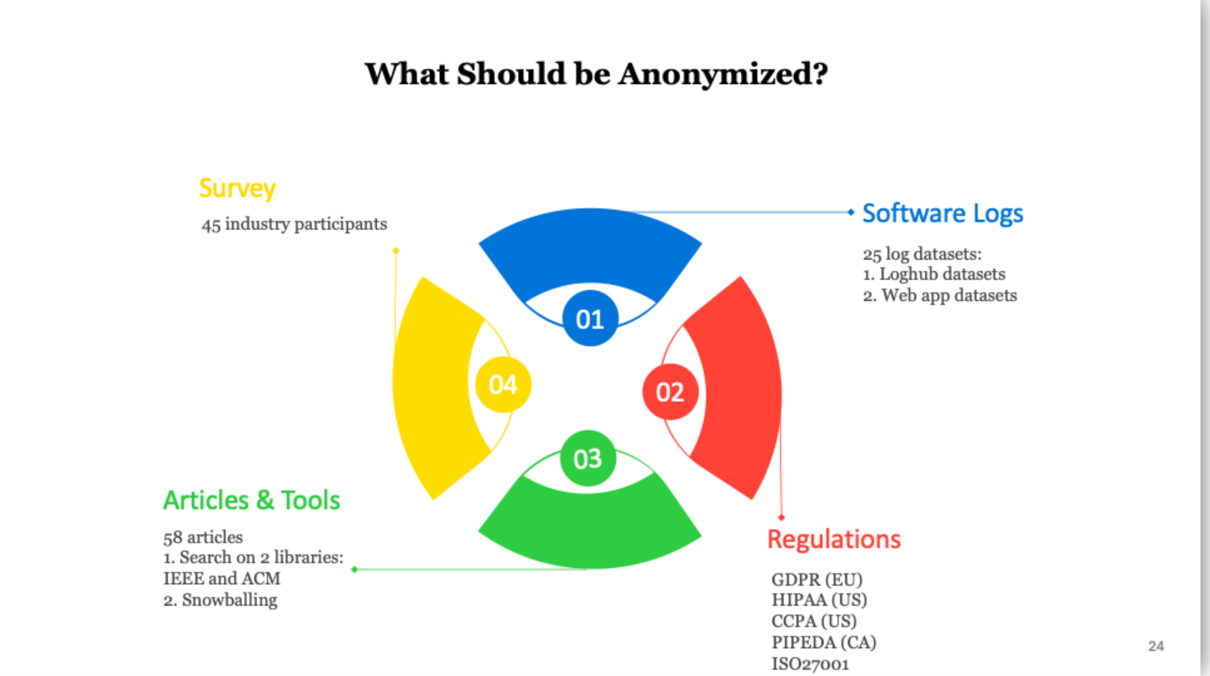


Personally Identifiable Information (PII)

- Name
- Email address
- Phone number
- Security number
- Driver's license number

Quasi-Identifiers

Date of birth
Gender
Postal code



What Should be Anonymized?

Survey

45 industry participants

Software Logs

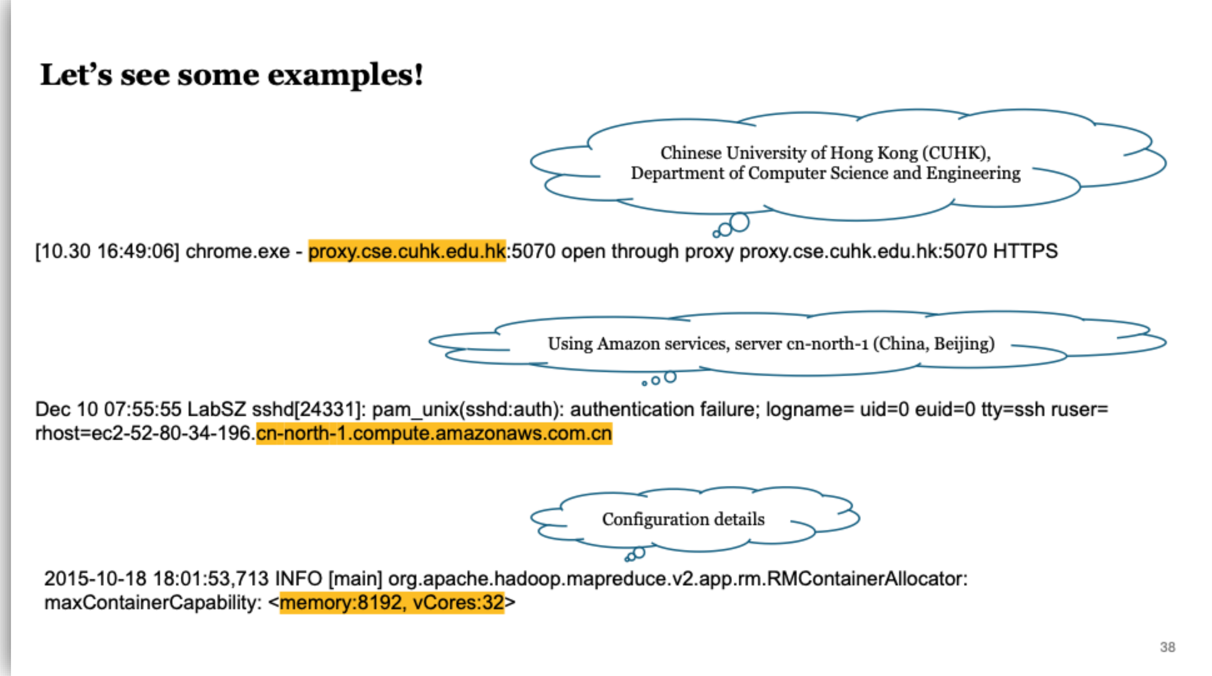
25 log datasets:
1. Loghub datasets
2. Web app datasets

Articles & Tools

58 articles
1. Search on 2 libraries: IEEE and ACM
2. Snowballing

Regulations

GDPR (EU)
HIPAA (US)
CCPA (US)
PIPEDA (CA)
ISO27001



Let's see some examples!

Chinese University of Hong Kong (CUHK), Department of Computer Science and Engineering

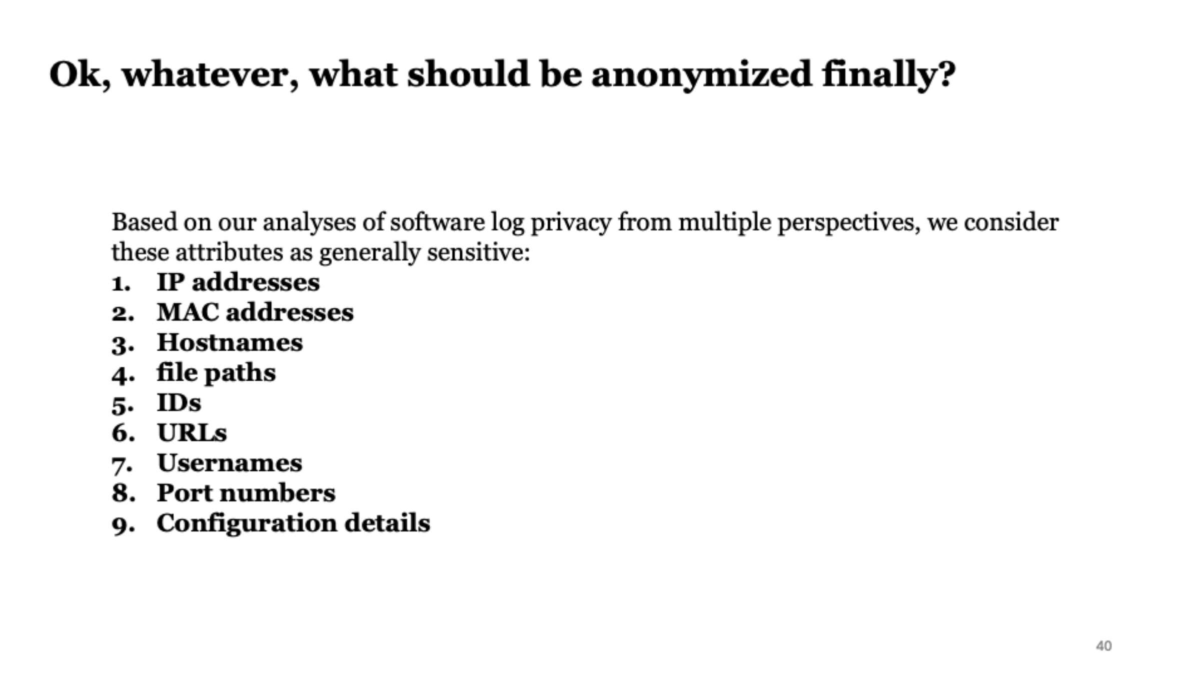
```
[10.30 16:49:06] chrome.exe - proxy.cse.cuhk.edu.hk:5070 open through proxy proxy.cse.cuhk.edu.hk:5070 HTTPS
```

Using Amazon services, server cn-north-1 (China, Beijing)

```
Dec 10 07:55:55 LabSZ sshd[24331]: pam_unix(sshd:auth): authentication failure; logname= uid=0 euid=0 tty=ssh ruser=
rhost=ec2-52-80-34-196.cn-north-1.compute.amazonaws.com.cn
```

Configuration details

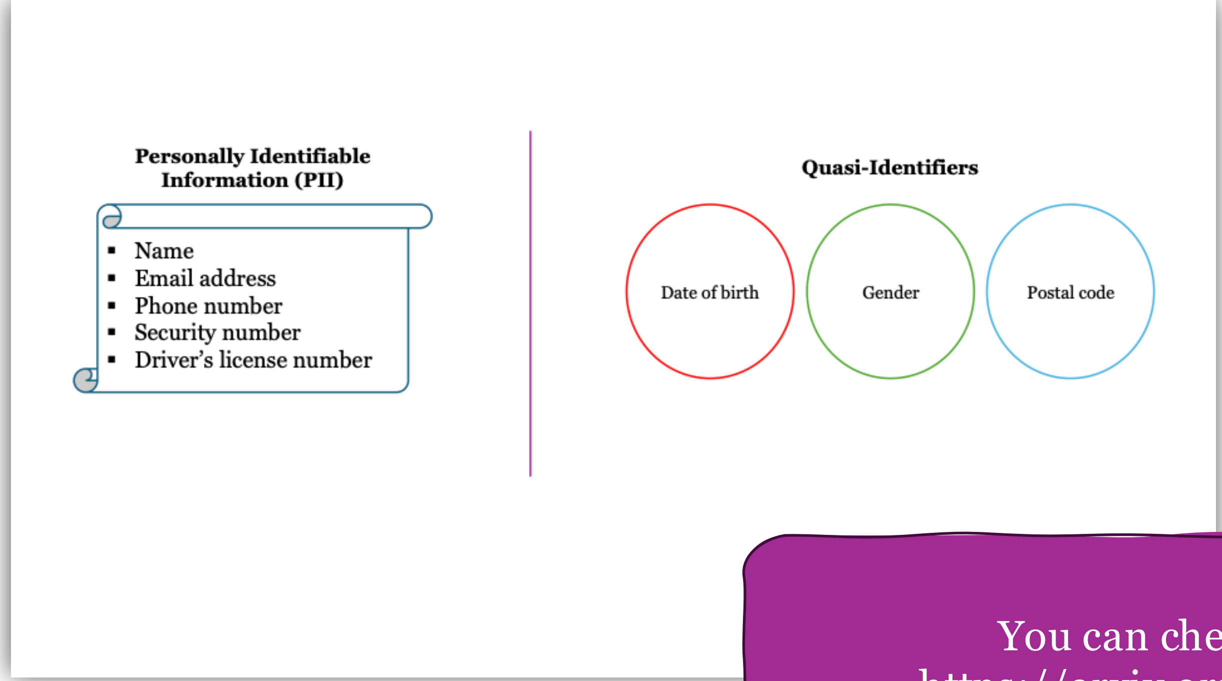
```
2015-10-18 18:01:53,713 INFO [main] org.apache.hadoop.mapreduce.v2.app.rm.RMContainerAllocator:
maxContainerCapability: <memory:8192, vCores:32>
```



Ok, whatever, what should be anonymized finally?

Based on our analyses of software log privacy from multiple perspectives, we consider these attributes as generally sensitive:

1. **IP addresses**
2. **MAC addresses**
3. **Hostnames**
4. **file paths**
5. **IDs**
6. **URLs**
7. **Usernames**
8. **Port numbers**
9. **Configuration details**



Personally Identifiable Information (PII)

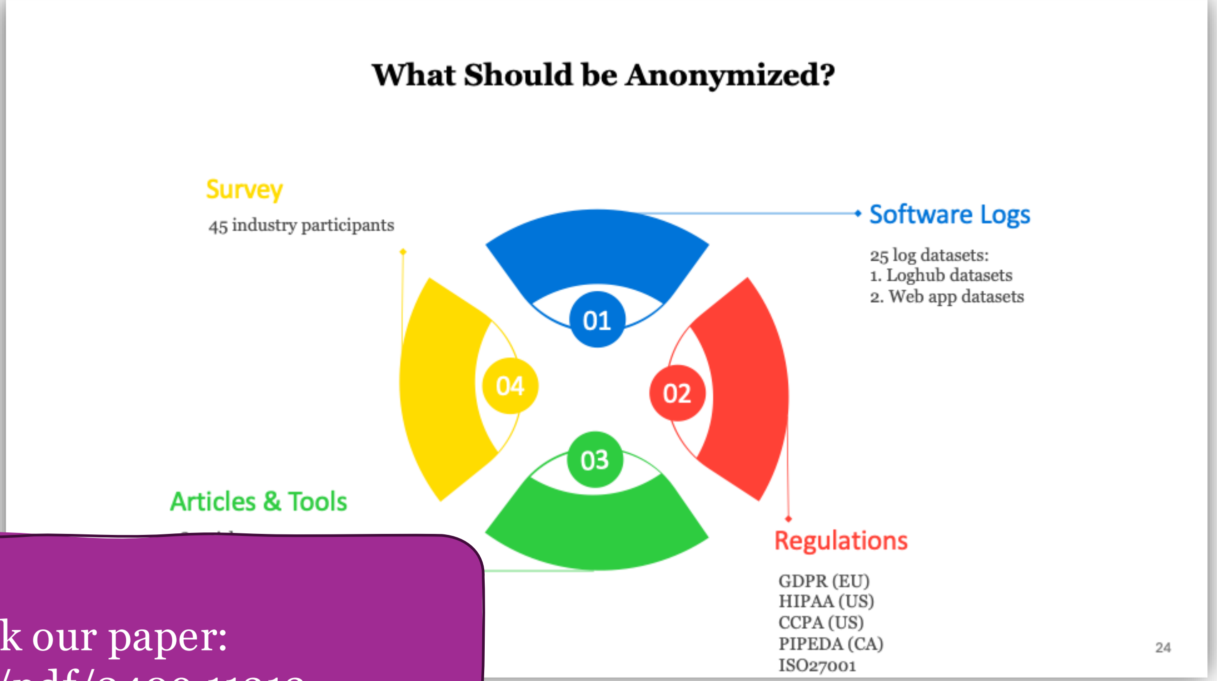
- Name
- Email address
- Phone number
- Security number
- Driver's license number

Quasi-Identifiers

Date of birth

Gender

Postal code



What Should be Anonymized?

Survey

45 industry participants

Software Logs

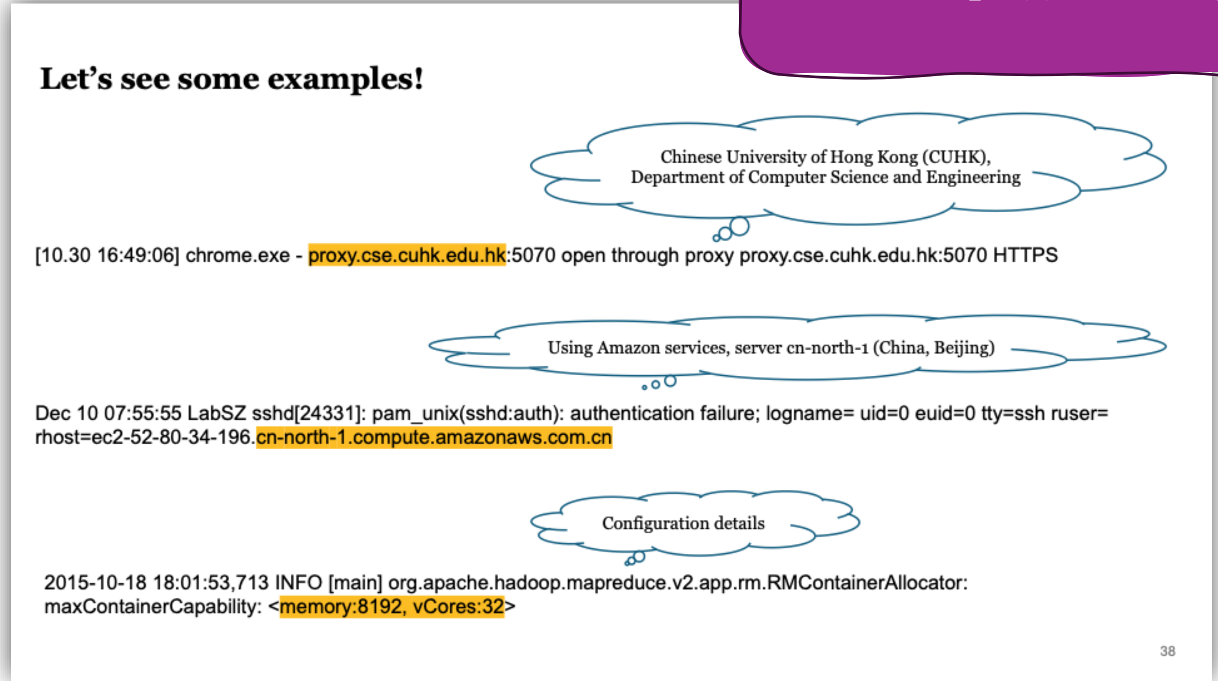
25 log datasets:
1. Loghub datasets
2. Web app datasets

Articles & Tools

Regulations

GDPR (EU)
HIPAA (US)
CCPA (US)
PIPEDA (CA)
ISO27001

You can check our paper:
https://arxiv.org/pdf/2409.11313



Let's see some examples!

Chinese University of Hong Kong (CUHK),
Department of Computer Science and Engineering

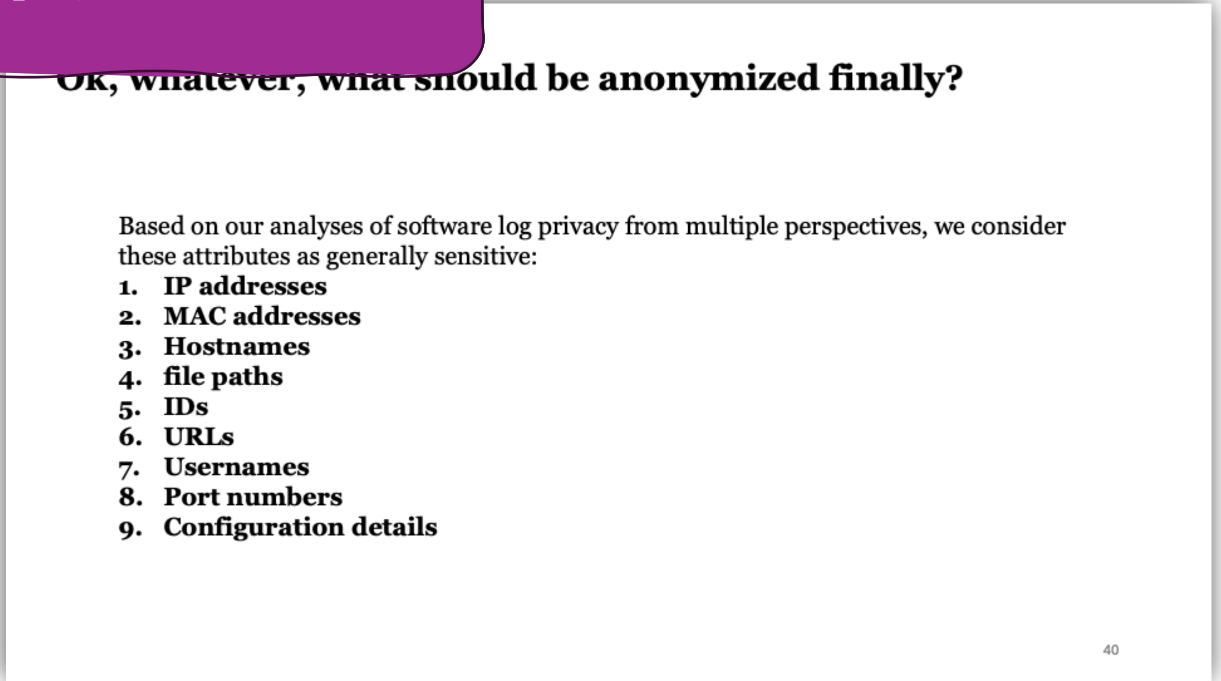
```
[10.30 16:49:06] chrome.exe - proxy.cse.cuhk.edu.hk:5070 open through proxy proxy.cse.cuhk.edu.hk:5070 HTTPS
```

Using Amazon services, server cn-north-1 (China, Beijing)

```
Dec 10 07:55:55 LabSZ sshd[24331]: pam_unix(sshd:auth): authentication failure; logname= uid=0 euid=0 tty=ssh ruser=
rhost=ec2-52-80-34-196.cn-north-1.compute.amazonaws.com.cn
```

Configuration details

```
2015-10-18 18:01:53,713 INFO [main] org.apache.hadoop.mapreduce.v2.app.rm.RMContainerAllocator:
maxContainerCapability: <memory:8192, vCores:32>
```



Ok, whatever, what should be anonymized finally?

Based on our analyses of software log privacy from multiple perspectives, we consider these attributes as generally sensitive:

1. **IP addresses**
2. **MAC addresses**
3. **Hostnames**
4. **file paths**
5. **IDs**
6. **URLs**
7. **Usernames**
8. **Port numbers**
9. **Configuration details**