Failure of Privacy Policy for Session Replay Services used for Monitor Your Keystroke

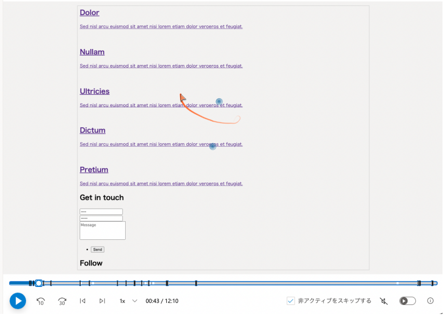
Daichi Kajima and Hiroaki Kikuchi

Abstract Session reply is a service used to capture the history of behavior performed in web browser. It is supposed to be used for detecting any error caused by human misunderstanding and improving the user experience. The captured data, including mouse movement events and typing keys, is classified as personal data that should be taken very carefully under the user's consent to be acquired. Privacy regulations such as GDPR requires user consent before obtaining personal identifiable information. However, we found that many major websites have session reply service without disclosing the use in their privacy policy. In this study, we have examined 11,523 domestic websites in Japan and clarify how many sites deploy hidden session replay service. Our analysis shows that 56 sites out of 300 sites have no description about the use of session replay service, that is, may be considered as illegal in terms of privacy regulation. After reporting the survey on session replay services, we propose a list of good practices to mitigate the failure of privacy policy.

1 Introduction

Session reply is a function to capture a visitor's journey on a web site in the form of history of events, keystrokes and mouse movements performed in the site. Session replay is supposed to help improving user experience and identifying obstacles in the interaction with. For example, Fig. 1 demonstrates that the user journey such as mouse movements and page scroll has been stored are replayed in Microsoft Clarity. It indicates where the users attention was paid in the web page and thus may be very useful to improve the user experience. According to our analysis (in Section 2), at least 981 sites activate session-replay service out of 11,523 Japanese website.

Daichi Kajima and Hiroaki Kikuchi
School of Interdisciplinary Mathematical Sciences, Meiji University, 4-21-1 Nakano, Tokyo, Japan
e-mail: kikn@meiji.ac.jp



Dolor

Sed nisl arcu euismod sit amet nisi lorem etiam dolor veroeros et feugiat.

Nullam

Sed nisl arcu euismod sit amet nisi lorem etiam dolor veroeros et feugiat.

Ultricies

Sed nisl arcu euismod sit amet nisi lorem etiam dolor veroeros et feugiat.

Dictum

Sed nisl arcu euismod sit amet nisi lorem etiam dolor veroeros et feugiat.

Pretium

Sed nisl arcu euismod sit amet nisi lorem etiam dolor veroeros et feugiat.

Get in touch

Message

Follow

Fig. 1 Microsoft Clarity service

The visitor's journey, however, contains personally identifiable information. For example, the keystroke conveys a unique feature of users. The journey may contain confidential credential such as passwords. Fig. 2 shows the actual keystroke captured and sent to the session replay server in Fullstory, where user name "kajima" was typed one by one. The example has an implication that confidential information may be disclosed to a third-party session-replay server without alerting it to visitor and even the web administrator who decide to set up the session replay service.

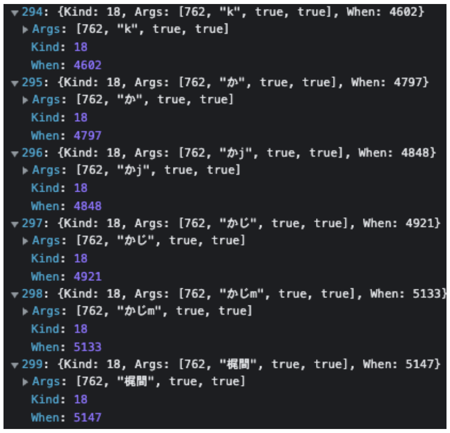
The potential risk of personal data disclosure is not explicit to visitor. The explicit way for alerting use of session-capture is a privacy policy. Unfortunately, many sites collect visitors' keystroke and mouse movement without specifying any statement in their privacy policy. According to our investigation (in Section 3), 56 our of 300 website (18.7%) have no description on the purpose of session capture or the name of session-replay services. We guess that most of them are innocent and do not notice that the session-replay service collect a personally identifiable information or a security credential.

The fact motivates us to do this study. Our research questions are as follows.

- How common web sites do set up session-replay service? What section of business are most popular for session-replay service?
- How frequenty a use of session-replay service is explicitly specified in a corresponding privacy policy? How about the the objectives of session-replay service?
- Which session-replay service are explicitly specified in a privacy policy?

To answer the these questions, we have conducted web site survey and the corresponding privacy policy in major top web sites in Japan. In paper, we aim to alart the risk of hidden capture of personally identified information and security credentials through widely deplored session-replay service.

The rest of this paper is organized as follows. In Section 2, we briefly describe the existing works related with session-replay services. We show the website survey for variety use of session-replay service in Section 3. The counts of usage brake up into several business sections. With the comparison of disclosure of the Google's Analytic, which is very known as one of the representative web usage statistics tool, we show that less the session-replay service are explicitly specified in a privacy



```
▼294: {Kind: 18, Args: [762, "k", true, true], When: 4602}
  ▶Args: [762, "k", true, true]
   Kind: 18
   When: 4602
▼295: {Kind: 18, Args: [762, "か", true, true], When: 4797}
  ▶Args: [762, "か", true, true]
   Kind: 18
   When: 4797
▼296: {Kind: 18, Args: [762, "かj", true, true], When: 4848}
  ▶Args: [762, "かj", true, true]
   Kind: 18
   When: 4848
▼297: {Kind: 18, Args: [762, "かじ", true, true], When: 4921}
  ▶Args: [762, "かじ", true, true]
   Kind: 18
   When: 4921
▼298: {Kind: 18, Args: [762, "かじm", true, true], When: 5133}
  ▶Args: [762, "かじm", true, true]
   Kind: 18
   When: 5133
▼299: {Kind: 18, Args: [762, "梶間", true, true], When: 5147}
  ▶Args: [762, "梶間", true, true]
   Kind: 18
   When: 5147
```

Fig. 2 Part of the transmission request at the time of text entry

policy. We propose some of measures toward disclosure the risks incurred from session-replay service. We conclude our study in Section 5.

2 Session-replay service

2.1 Related works

In 2020, Gunes et. al [2] conducted the large-scale experiment using open-source software OpenWPM to disclose the third-party script embedded in 50,000 websites. They disclose that the disclosure of personal data and security credentials stored in a DOM used in session-replay service, FullStory, UserReplay, SessionCam, Hotjar, Yandex and Smartlook.

In 2022, Xiufen et. al [3] investigated 19,483 hospital websites in Asia, North America, South America, Africa, Oceania for the security and privacy. They found that more than 690 websites have session-replay service including Hotjar, Yandex and FullStory and have sent the data, name, email address and password.

2.2 Preliminary survey

Which session replay services are most frequently used? We investigate the statistics of search queries that are related with session-replay services. Table 1 shows the statistics of top session-replay services in the number of queries used in the HTML contents. Our survey employees the *PublicWWW*[5] that provides the web page in conjunction with the HTML source codes for a given query. According to the sur-

Table 1 Search results in PublicWWW

service	domain	number of queries
Microsoft Clarity	clarity.ms	91,110
Hotjar	hotjar.com	293,807
mouseflow	mouseflow.com	27,641
crazyegg	crazyegg.com	28,186
Contentsquare	contentsquare	1,670
lucky orange	luckyorange.com	13,642
fullstory	fullstory.com	21,132
Yandex	yandex.ru	1,000,000
Dynatrace	dynatrace.com	6,924
Glassbox	glassbox	806
Smartlook	smartlook.com	33,014
Foresee	foresee.com	254
Inspectlet	inspectlet.com	10,340
LogRocket	logrocket	5,997

vey, we target the top 15 services; Microsoft Clarity[6], Hotjar[7], mouseflow[8], crazyegg[9], Contentsquare[10], lucky orange[11], fullstory[12], Yandex[13], Dynatrace[14], Glassbox[15], Smartlook[16], foresee[17], Inspectlet[18], and LogRocket[19].

2.3 Survey of deployments of session replay

The deployment of session-replay service depends on a kind of business sectors. We investigate the top 11,523 web sites ranked in the Tranco [1] over the major 15 session-replay services. We use the Selenium[4] framework for automated retrieval the list of HTTP requests sent from the given web site. For instance, a web site sends HTTP request something like `https://www.clarity.ms/tag/7vgto77nxr?ref=gtm`, for which we detect the Microsoft Clarity is set up in the site.

Table 2 shows the result of our investigation. We detect 981 session-replay-activated sites (8.51 %) out of 11,523 web sites. The top service is Microsoft Clarity (702, 6.1 %). The most advanced sections for session-replay include information and communication service (ICT), wholesale trade. A possible reason why most site in these sections set up session-replay is that these business often provide online shops for which monitoring visitor behaviors are very useful.

3 Insufficient Privacy Policies

3.1 Objectives

How often companies disclose a use of session-replay service in the privacy policy explicitly? We investigate 197 websites (manually) chosen from each of business

Table 2 Number of installations per service at domestic sites, (MR (mouseflow), CS (Contentsquare), CE (crazyegg), DT (Dynatrace), FS (Foresee), GB (glassbox), IL (inspectlet), LR (logrocket), LR (luckyorange), SL (smartlook)

sectors	Clarity	Hotjar	MR	Yandex	CS	CE	DT	FS	fullstory	GB	IL	LR	LR	SL	total
ICT	274	33	22	3	2	6	0	0	3	1	0	0	1	1	346
wholesale	127	19	9	0	11	19	0	1	1	0	0	0	4	1	192
industory	40	13	5	0	4	6	2	0	1	0	0	0	1	0	72
hotel	21	6	1	0	1	0	0	0	0	1	0	0	0	0	30
educational	18	5	4	0	0	1	0	0	1	0	0	0	0	0	29
entertainment	20	3	2	0	0	0	0	0	0	0	0	0	0	0	25
academic study	34	2	3	0	0	1	0	0	0	0	0	0	0	0	40
general service	76	3	8	0	0	3	0	0	0	0	0	0	1	0	91
financial business	13	1	3	0	2	1	0	0	0	0	0	0	0	0	20
construction	7	0	3	0	0	0	0	0	0	0	0	0	0	0	10
electric power	4	0	1	0	0	0	0	0	0	0	0	0	0	0	5
medical service	6	2	1	0	0	0	0	0	0	0	0	0	0	0	9
real estate	15	0	5	1	0	0	0	0	0	0	0	0	0	0	21
transportation	7	0	1	0	0	0	0	0	0	0	0	0	0	0	8
integration	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
unknown	39	2	0	0	0	3	0	0	0	0	0	0	0	0	44
total	702	89	68	4	20	40	2	1	6	2	0	0	7	2	943

Table 3 Session-replay-disclosure statements in the privacy policy

service	N	both	objective	service name	neither
Microsoft Clarity	50	12	35	0	3
Hotjar	50	6	41	0	3
mouseflow	50	2	29	0	19
crazyegg	50	10	31	0	9
ContentSquare	26	2	22	0	3
lucky orange	20	0	14	0	6
fullstory	6	1	2	0	1
Yandex	15	1	10	0	4
Dynatrace	8	0	8	0	0
Glassbox	4	2	2	0	0
Smartlook	14	0	7	0	7
Foresee	2	0	2	0	0
Inspectlet	3	0	3	0	0
LogRocket	2	1	0	0	1
total	300	37	207	0	56

sectors (1) whether the name of session-replay service is published, and (2) whether the objective of session-replay service is published.

Table 3 shows the survey results. We found that 37 out of 300 (12.3 %) web sites that disclose both of the objective of session capture and the name of session-replay service used, and 56 (18.7%) sites has no description at all.

Table 4 Google Analytics disclosure Statement in the privacy policy

service	*N*	deploy	disclosure
Microsoft Clarity	50	49	33
Hotjar	50	45	12
mouseflow	50	46	11
crazyegg	50	46	21
ContentSquare	27	23	13
lucky orange	20	20	3
fullstory	6	6	3
Yandex	15	14	4
Dynatrace	8	7	6
Glassbox	4	3	3
Smartlook	15	14	2
Foresee	2	2	1
Inspectlet	3	3	2
LogRocket	2	2	2
total	297	276	115

3.2 Comparison with disclosure of Google Analytics

The disclosing rate of 12.3% is quite low in comparison with the disclosure of the Google Analytics, well-known web site monitoring service.

Table 4 shows the statistics of disclosure of Google Analytics for the same web sites that employ major session-replay services. Most of web sites using any of session-replay service also deploys Google Analytics. The fraction of both services is 277 (92.9%). In contrast, 115 sites disclose the deployment of Google Analytics (41.7which is quite higher than that of session-reply disclosure. We find that Google requires the disclosure of its service. Hence, it helps to increase the deployment rate.

4 Mitigation

Based on our analysis, we would like to propose some of mitigation to dismiss the inconsistency between the risk of session-replay and the reception of visitors. We believe that both the service providers and the users need to go forward for consensus of session-replay service.

1. Service providers should disclose the list of acquired personally identifiable information from the web site in their privacy policy explicitly.
2. Service providers should obtain visitor's consent for acquirement of session related information including mouse movements and keystrokes.
3. Users should notice the privacy policy statement before they visit web sites.
4. Users should avoid unnecessary providing personally unidentifiable information through web site.

5 Conclusions

We have studies the risk of session-replay service for sending keystroke without noticing. Our investigation for major 15 session-replay derives in 11,523 web sites in Japan showed that only few site (11.6%) disclose the name of session-replay service from their privacy policy, while 981 web sites (8.51%) deploy the session-replay service. Based on the potential risk of unauthorized personal data acquirement, we proposed some of mitigation practices to dismiss the acquirement of keystroke without visitor's consent. We plan to conduct a comprehensive survey of deployment of session-replay service.

References

1. Tranco, "Tranco A Research-Oriented Top Sites Ranking Hardened Against Manipulation" (https://tranco-list.eu/, December 2022).
2. Gunes Acar, Steven Englehardt, Arvind Narayanan, "No boundaries: data exfiltration by third parties embedded on web pages", Proceedings of the 20th Privacy Enhancing Technologies Symposium (PETS), Pages 1-19, 2020.
3. Xiufen Yu, Nayanamana Samarasinghe, Mohammad Mannan, Amr Youssef, "Got Sick and Tracked: Privacy Analysis of Hospital Websites", IEEE European Symposium on Security and Privacy Workshops, pp 278-286, 2022.
4. Selenium, "The Selenium Browser Automation Project", (https://www.selenium.dev/documentation/, accessed in Feb. 2023).
5. PublicWWW, (https://publicwww.com/, accessed in Feb. 2023).
6. Microsoft Clarity, (https://clarity.microsoft.com/, accessed in Feb. 2023).
7. Hotjar, (https://www.hotjar.com/, accessed in Feb. 2023).
8. Mouseflow, (https://mouseflow-jp.com/, accessed in Feb. 2023).
9. Crazyegg, (https://www.crazyegg.com/, accessed in Feb. 2023).
10. Contentsquare, (https://contentsquare.com/jp-jp, accessed in Feb. 2023).
11. lucky orange, (https://www.luckyorange.com, accessed in Feb. 2023).
12. fullstory, (https://www.fullstory.com, accessed in Feb. 2023).
13. Yandex, (https://metrica.yandex.com/about, accessed in Feb. 2023).
14. Dynatrace, (https://www.dynatrace.com/ja, accessed in Feb. 2023).
15. Glassbox, (https://www.glassbox.com, accessed in Feb. 2023).
16. Smartlook, (https://www.smartlook.com, accessed in Feb. 2023).
17. Foresee, (https://www.verint.com, accessed in Feb. 2023).
18. Inspectlet, (https://www.inspectlet.com, accessed in Feb. 2023).
19. LogRocket, (https://logrocket.com, accessed in Feb. 2023).
20. Google Analytics, (https://marketingplatform.google.com/intl/ja/about/analytics, accessed in Feb. 2023).