Tracing and Analyzing Web Access Paths based on User-Side Data Collection: How Do Users Reach Malicious URLs?

Takeshi Takahashi, Ph.D.,
NICT

Most of the contents in this talk was already published in the following publication:

T.Takahashi, C.Kruegel, G.Vigna, K.Yoshioka, D.Inoue, "Tracing and Analyzing Web Access Paths Based on User-Side Data Collection: How Do Users Reach Malicious URLs?," RAID 2020

Additional analysis results will be also presented in this talk.

Agenda

1. Background
2. Dataset
3. Hazardous paths reconstruction scheme
4. Analysis of the first accesses of the hazardous paths
5. Preemptive domain filtering scheme

Background

- Web access is a major channel of malware infection
- Many techniques have been considered to cope with this issue
  - Blacklist /whitelist
  - Machine-learning-based assessments
- However, there are still users who get accesses to these sites.
- To better protect users, we need to know how users reach these malicious web sites by collecting large-scale web access records and by analyzing them in detail

In this work, we collect users' access records and analyze their access paths to better protect users.

1. We collect users' access records through our browser extension.
2. We reconstruct the access path by analyzing the records.
3. We analyze the entry points of hazardous paths, which are the paths leading to malicious URLs
4. We introduce our preemptive domain filtering scheme, which identifies domains that often lead to malicious URLs.

Agenda

1. Background
2. Dataset
3. Hazardous paths reconstruction scheme
4. Analysis of the first accesses of the hazardous paths
5. Preemptive domain filtering scheme

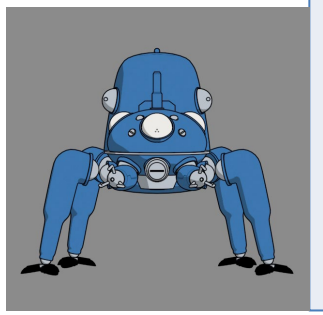
Per-user data collection on the browser

Approach

- We prepared a browser extension that runs on Chrome browser
- To motivate people to keep using the extension after its installation, we used a popular character, called “Tachikoma”
- It records each user’s access log and shares the information to our server periodically.
- The server sends the collected URLs to the GSB once a day and store the evaluation results internally.

Ethical Considerations

- We worked with our Internal Review Board to ensure that our usage of the logs was ethical and respectful of users’ privacy.
- See the paper for the details.

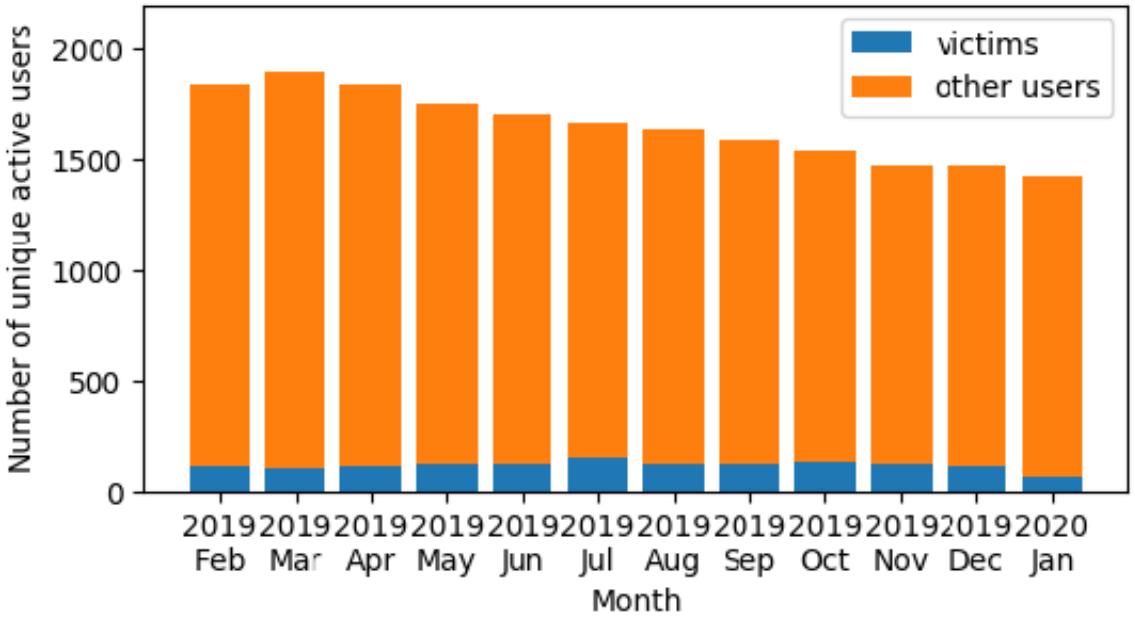


List of information

- User ID
- Tab ID
- URL
- Tab URL
- Source tab ID
- Timestamp
- Referer
- Resource type (main frame, sub frame, etc)
- Transition type (auto_bookmark, link, etc)
- GSB evaluation results

Overview of our dataset

The number of active users



Collection period: Feb. 1, 2019—Jan. 31, 2020

Access records: 4,306,529,287 access records, among which 76,474 records are the access records to blacklisted URLs

Number of users: 1,650 users per month, among which 115 users reached malicious URLs

Agenda

1. Background
2. Dataset
3. Hazardous paths reconstruction scheme
4. Analysis of the first accesses of the hazardous paths
5. Preemptive domain filtering scheme

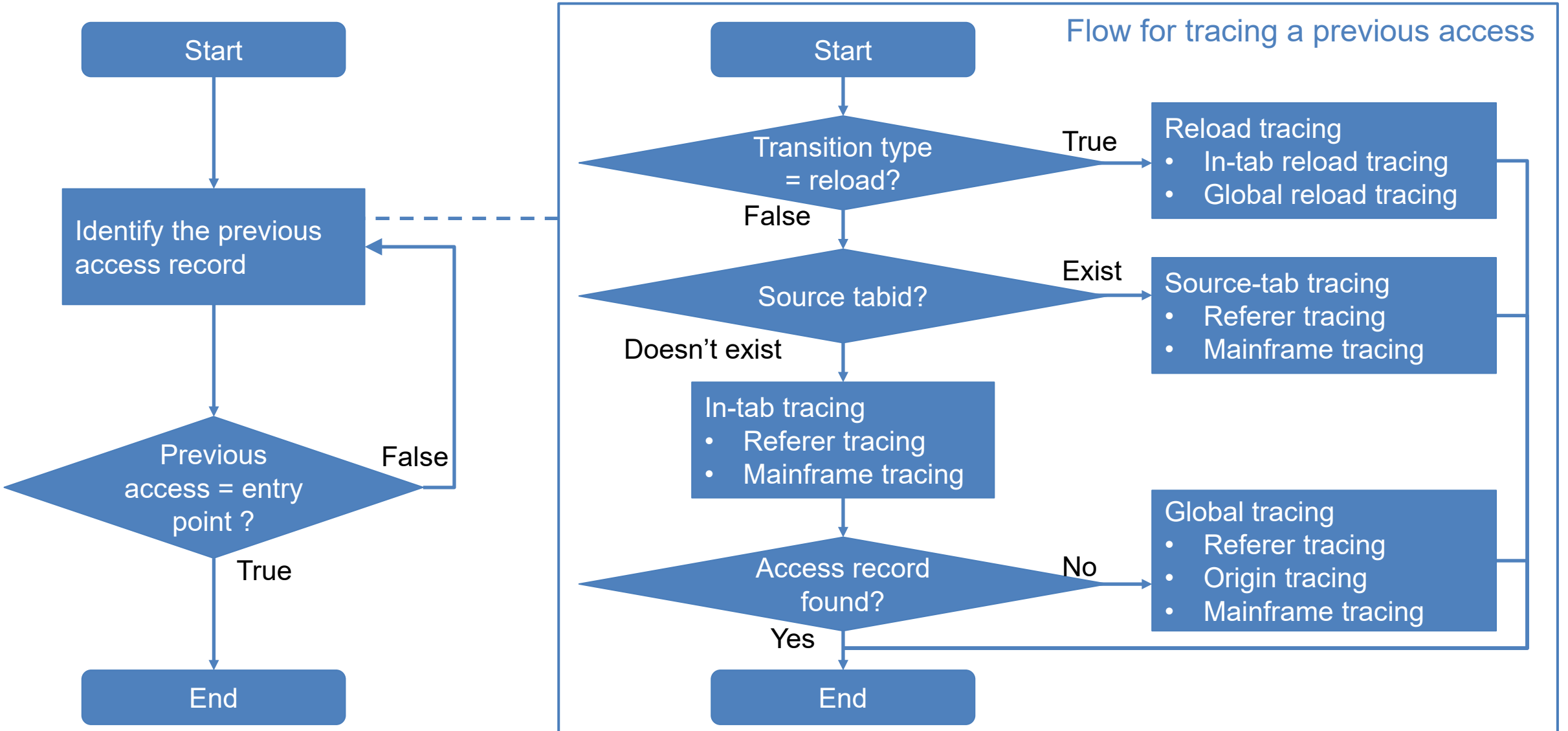
Method for reconstructing a hazardous path

- We iteratively trace previous accesses until we reach an entry point

- Entry point is discontinuous from its previous access, including the following types of accesses
  - Bookmark access
  - Session reconstruction
  - Web search
  - Omnibar access
  - Address typing
  - Start page access

- We minimize the range of logs that we need to analyze by identifying user IDs and tab IDs
  - We minimize the ambiguity of log tracing
  - We minimize the time, resource, and cost of analysis

Method for reconstructing a hazardous path



An example of a reconstructed hazardous path

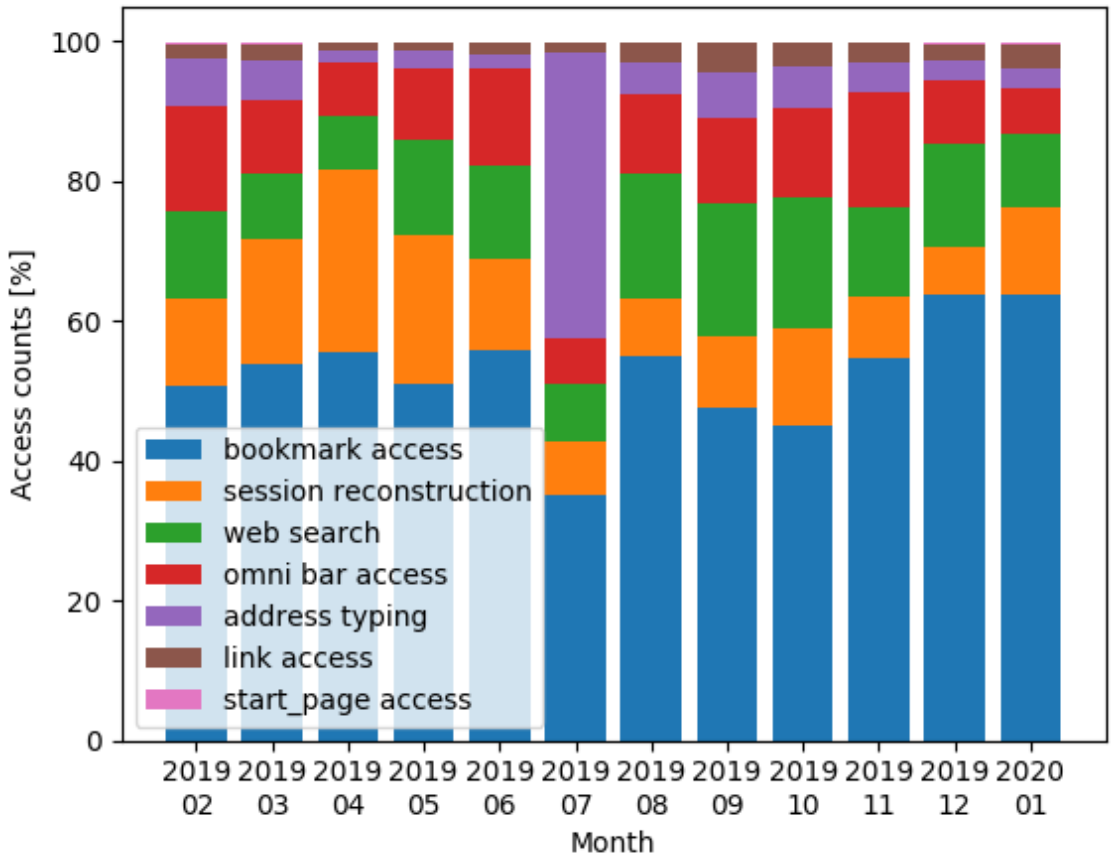
Time (JST)	Tab ID	URL (excerpt)	Source tab ID	Transition type	Resource type	Tracing method
15:26:57	182	http://javtorrent.re/category/		auto bookmark	main frame	In-tab tracing
15:27:14	182	http://javtorrent.re/?s=080819		form submit	main frame	In-tab tracing
15:27:26	182	http://javtorrent.re/uncensore		link	main frame	In-tab tracing
15:28:28	182	http://javtorrent.re/?s=HEYZO-		form_submit	main frame	In-tab tracing
(omitted 18 access records)						
15:31:50	182	http://javtorrent.re/uncensore		link	main frame	Source tab tracing
15:32:08	403	https://www.google.com/search?	182	Link	main frame	In-tab tracing
15:32:35	403	https://7mmtv.tv/zh/uncensored		Link	main frame	In-tab tracing
15:33:26	403	https://www.google.com/search?		Link	main frame	In-tab tracing
15:33:37	403	http://javhuge.com/Momoki%20		---	Complemented frame	In-tab tracing
15:33:37	403	http://javhuge.com/zb users/th		---	Style sheet	---

Agenda

1. Background
2. Dataset
3. Hazardous paths reconstruction scheme
4. Analysis of the first accesses of the hazardous paths
5. Preemptive domain filtering scheme

Breakdown of the entry point types of hazardous paths

Types	On hazardous paths		On all paths	
Bookmark access	4,062	(50.6%)	1,966,881	(38.1%)
Web search	1,168	(14.5%)	840,709	(16.3%)
Session reconstruction	985	(12.3%)	934,307	(18.1%)
Omni bar access	789	(9.8%)	835,206	(16.2%)
Address typing	699	(8.7%)	131,794	(2.6%)
Start page access	237	(3.0%)	448,912	(8.7%)
Link access	94	(1.2%)	---	
Total	8,034	(100%)	5,157,809	(100%)



Typical scenario:

- Access to a bookmarked page, e.g., link collection page or bulletin board, which lead to malicious URLs.
- Search the web to obtain the product code, e.g., that of porn video, and re-search the web with the product code, accessing illegal pages that are often malicious.

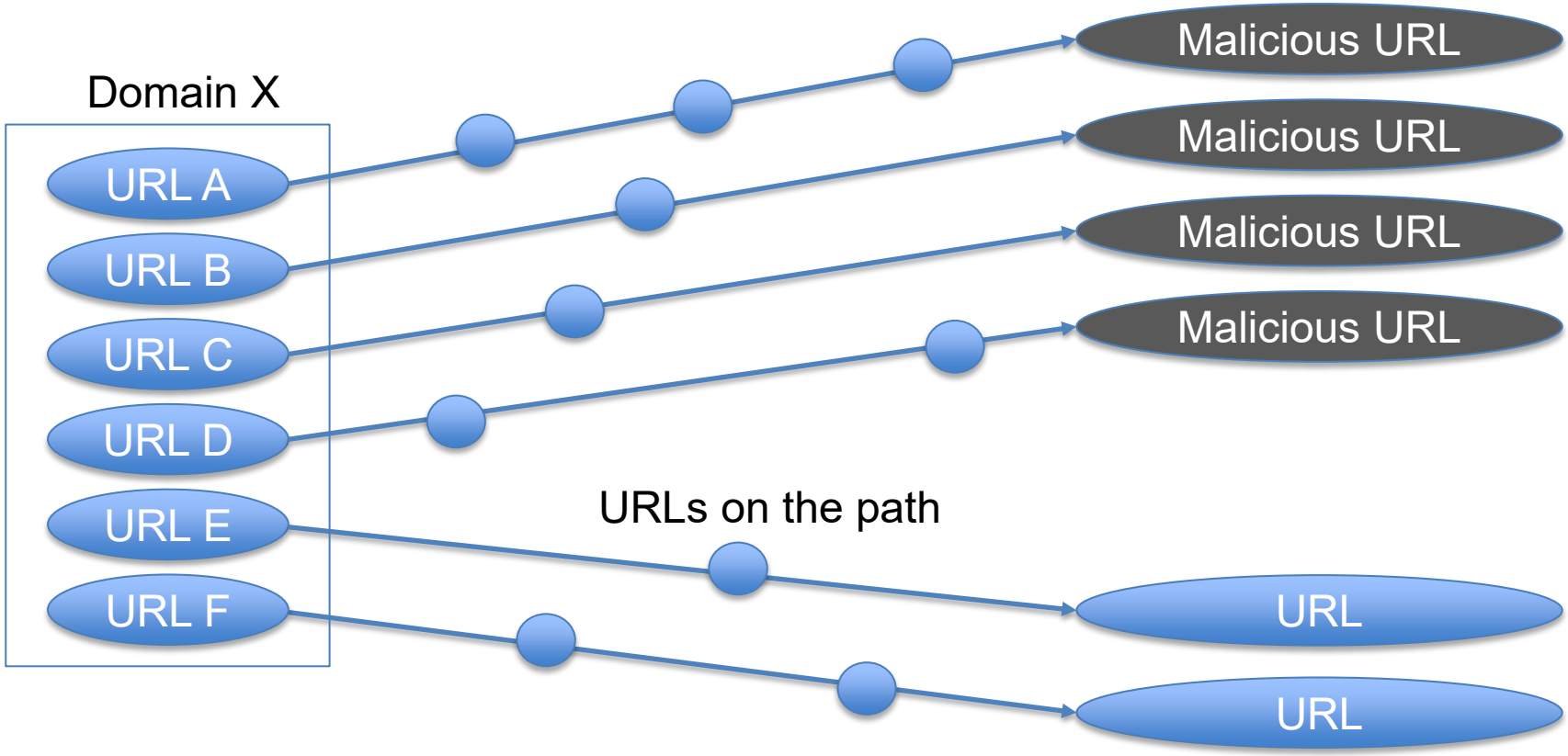
Agenda

1. Background
2. Dataset
3. Hazardous paths reconstruction scheme
4. Analysis of the first accesses of the hazardous paths
5. Preemptive domain filtering scheme

Domain risk level calculation

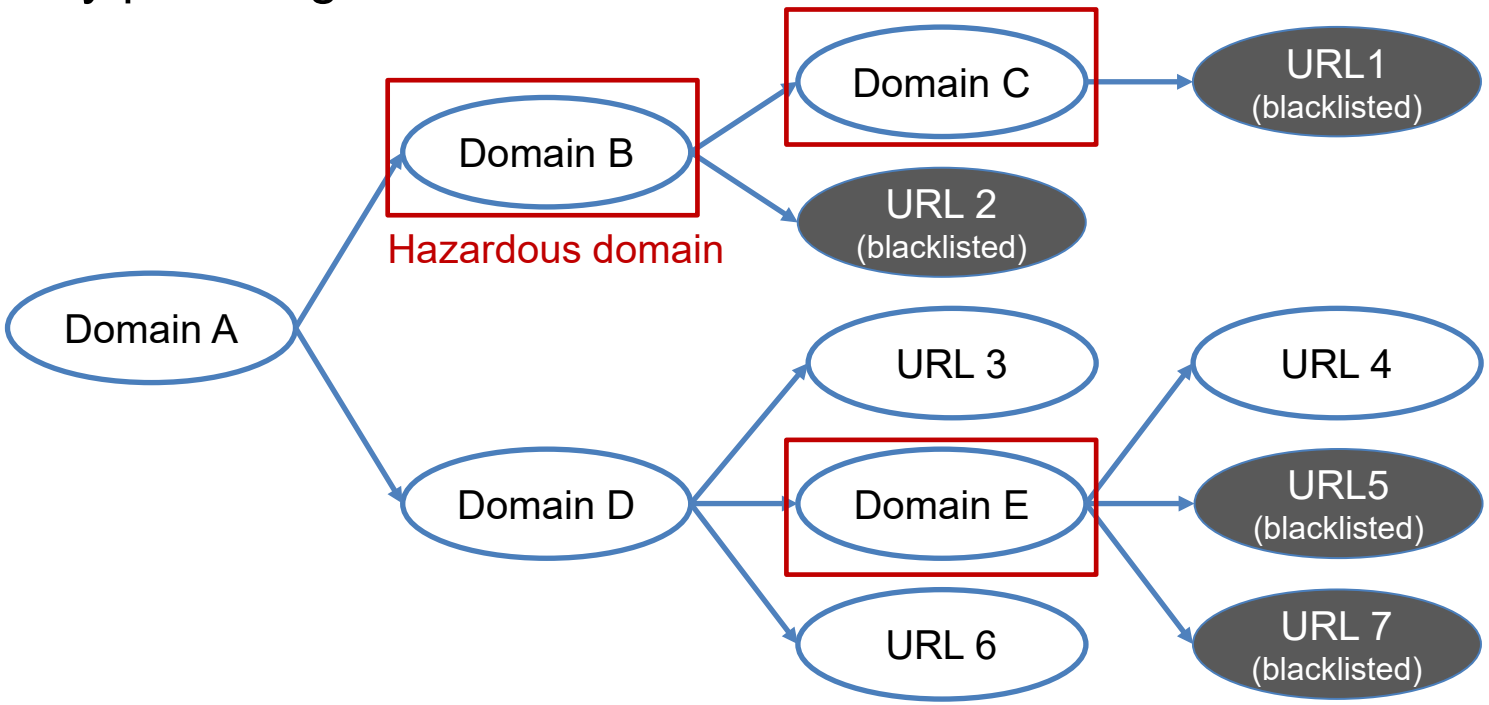
A domain risk level is defined as the probability of reaching a malicious URL after visiting a domain. It is the frequency of reaching malicious URLs divided by the number of accesses to the domain.

Example: Domain X is accessed 6 times, among which 4 accesses reach malicious URLs eventually. In this case, the risk level is 4/6, i.e., 66.7%.



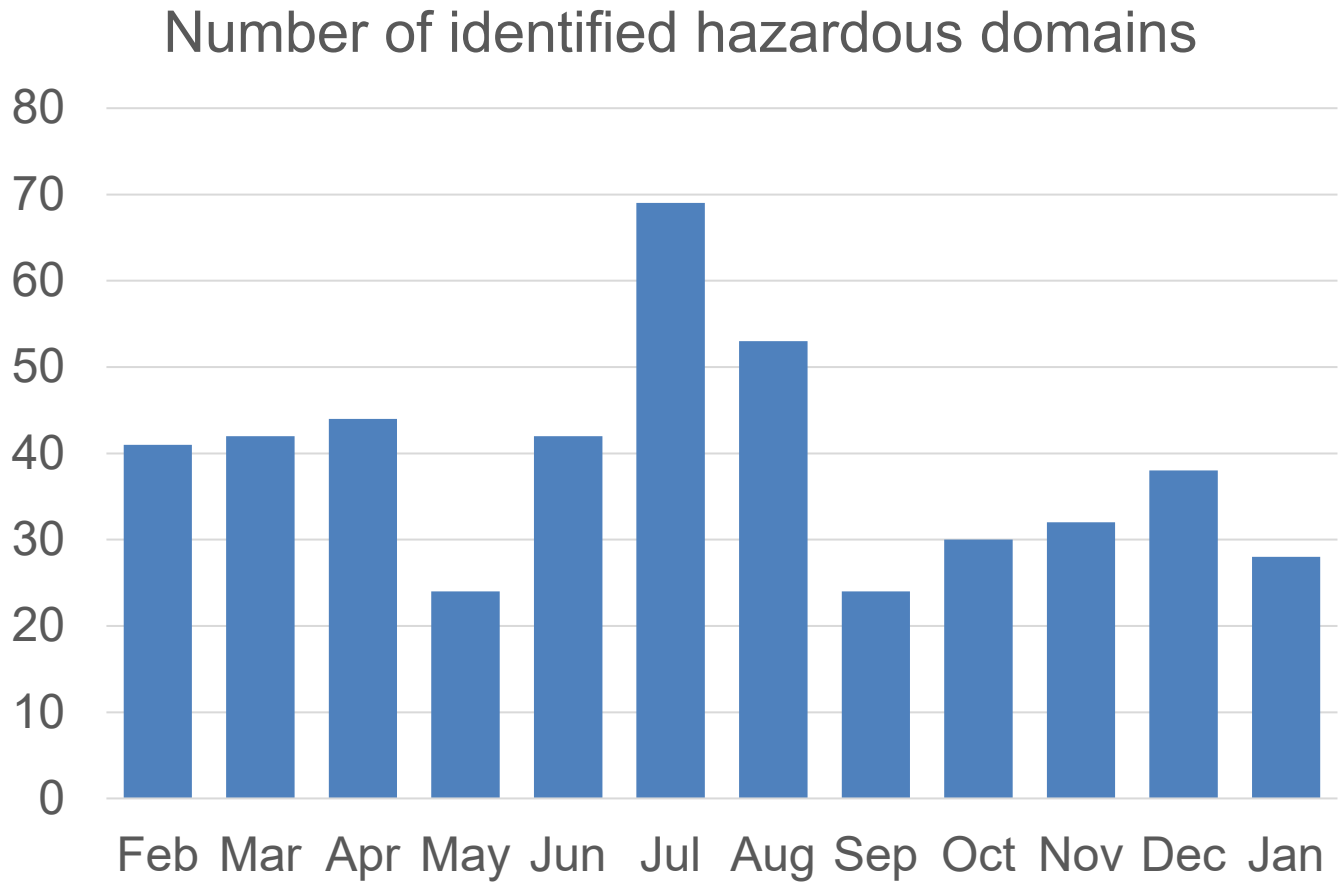
Domain risk evaluation method

- Different from schemes identifying malicious URLs and domains, the proposed scheme identifies domains that most likely lead to malicious URLs.
- These domains themselves do not necessarily host malicious contents, thus many of these are not listed on blacklists
- Risk levels of all domains on the paths are calculated, and those domains with the risk levels beyond certain threshold value are determined as hazardous
- We can minimize the number of users and accesses reaching malicious URLs by filtering traffic on the domains or by providing alerts.



Distribution of domain risk levels, monthly

Risk level	Is the domain registered by GSB?			Total
	Yes	Partially	None	
100%	903	15	467	1,385
[80%-100%)	15	0	29	44
[60%-80%)	22	0	84	106
Total	940	15	580	1,535



We have identified 467 domains that are not listed by GSB and that surely lead to malicious URLs.

URLs that become unreachable

URLs that become unreachable by the traffic blocking at the hazardous domains fall within one of the followings:

1. URLs listed on GSB
2. URLs listed on the other blacklists
3. Non-blacklisted URLs belonging to the same domain as those blacklisted URLs
4. Unreachable URLs
5. URLs with illegitimate or harmful contents (these are not listed by the Alexa Top 1,000 sites)

Blocking access to these URLs would help improving user protection without impairing users' legitimate activities.

Summary and conclusions

- We have reconstructed and analyzed hazardous paths by collecting access records at the user side, revealing that
  - bookmark access is the largest entry point to hazardous paths
  - its share is larger on hazardous paths than any other paths.

- We have proposed preemptive domain filtering scheme
  - It identifies domains that most likely lead to malicious URLs even if these domains do not host any malicious contents
  - It enables us to block traffic on these domains or provide alerts.
  - We have identified 467 domains that are not listed by GSB and that surely lead to malicious URLs.

Thank you very much for your kind attention.

Contact: takeshi_takahashi@ieee.org