FP-Stalker:

Tracking Browser Fingerprint Evolutions

Antoine Vastel, Pierre Laperdrix, Walter Rudametkin, Romain Rouvoy

Wednesday 23rd May, 2018

University of Lille / Inria, France

Browser Fingerprinting: Stateless Tracking

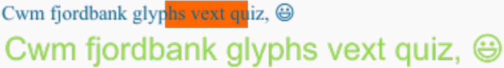
Objective: Track users over multiple visits

- Especially useful when deleting cookies

Approach: Load an extra script that:

- Generates a unique identifier from a device configuration
- Exploits the diversity of configurations

Example of a Browser Fingerprint

Attribute	Value
`Encoding`	gzip, deflate, sdch, br
`Languages`	en-US,en;q=0.8,es;q=0.6
`User-agent`	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/54.0.2840.99 Safari/537.36
`Canvas`	 Cwm fjordbank glyphs vext quiz, 😃 Cwm fjordbank glyphs vext quiz, 😃
`Platform`	Win32
`Resolution`	2560x1440x24

Related Work

Fingerprint uniqueness: 80–90 % [PETS 2010, S&P 2016]

But uniqueness is not enough for tracking: we also need stability [WWW 2015]

Objectives of this paper:

1. Evaluate fingerprint stability
2. Evaluate the effectiveness of browser fingerprint tracking

Amiunique dataset

https://amiunique.org:

- 1 website
- 2 browser extensions (Chrome and Firefox)

2 years: From July 2015 to early August 2017

98,598 fingerprints gathered from 1,905 distinct browsers (data cleaned)

Fingerprint stability

Stability varies depending on the attribute and the user

	Percentile (days)		
Attribute	**50th**	**90th**	**95th**
Resolution	Never	3.1	1.8
User agent	39.7	13.0	8.4
Canvas	290.0	35.3	17.2
Language	Never	215.1	56.7
Accept	Never	163.8	109.5
Cookies	Never	Never	Never

Tracking definition

Definition: Tracking is the process of linking fingerprints from a given browser

2 options:

1. Identical/similar fingerprint: link to an existing browser identifier
2. No/too many similar fingerprints: assign a new browser identifier

Rule-based linking algorithm

Strict rules:

- OS, platform and browser family must be identical
- Browser version is constant or increasing

Statistical rules:

- Local storage, . . . , canvas ⇒ must be identical
- Similarity of User agent, ..., headers ⇒ must be > 0.75
- Resolution, timezone can be different
- No more than 2 attribute changes

Hybrid approach: Rules + Machine learning

Our hybrid approach combines:

1. Rules: Use strict rules to filter candidates
2. Machine learning: Apply supervised ML to increase accuracy

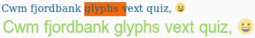
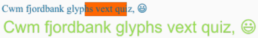
Machine learning model

Compute the probability that 2 fingerprints originate from the same browser

Random forest:

- Multiple decision trees
- Vote between different decision trees
- Tradeoff between precision and interpretability

Vectorization of fingerprints

Attribute	FP new	FP database	Vector
Encoding	”gzip, deflate, br”	”gzip, deflate”	0.87
Languages	”en-US,en;q=0.5”	”fr-FR,fr;q=0.8,en-US; q=0.6,en;q=0.4”	0.53
Canvas	 Cwm fjordbank glyphs vext quiz, 😀 Cwm fjordbank glyphs vext quiz, 😀	 Cwm fjordbank glyphs vext quiz, 😀 Cwm fjordbank glyphs vext quiz, 😀	0
...	...	...	...
Number changes			4

Training phase

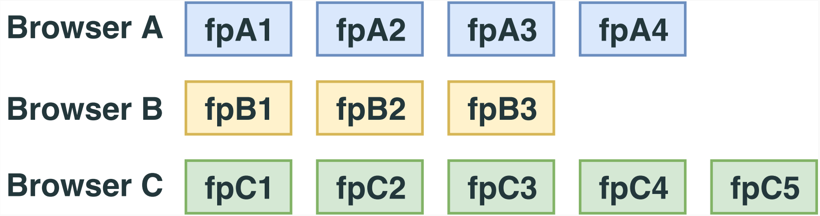
Train the random forest model:

- Training set composed of 40 % data chronologically ordered
- Feed pairs of fingerprints to the algorithm
- Apply undersampling to reduce overfitting

Evaluation

Evaluate the effectiveness of browser fingerprint tracking

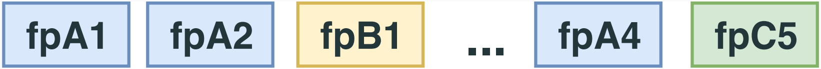
Test set: 59, 159 fingerprints from 1, 395 browsers



Browser A	fpA1	fpA2	fpA3	fpA4	
Browser B	fpB1	fpB2	fpB3		
Browser C	fpC1	fpC2	fpC3	fpC4	fpC5

Generate fingerprint sequence

Simulate the fingerprinting frequency (1 day, 2 days, ..., 20 days)

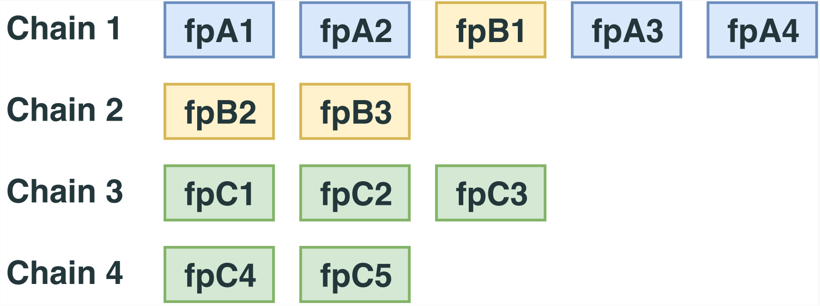


fpA1 fpA2 fpB1 ... fpA4 fpC5

Goal: compare tracking effectiveness at different collect frequencies

Apply linking algorithms

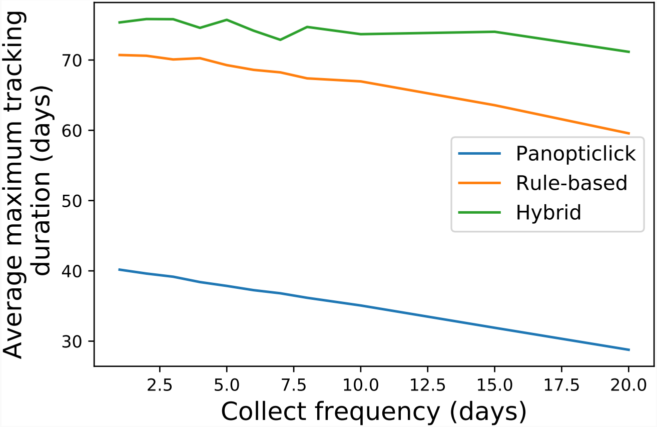
Link each fingerprint in the generated test set (chronologically)



Chain 1	fpA1	fpA2	fpB1	fpA3	fpA4
Chain 2	fpB2	fpB3			
Chain 3	fpC1	fpC2	fpC3		
Chain 4	fpC4	fpC5			

Average maximum tracking duration

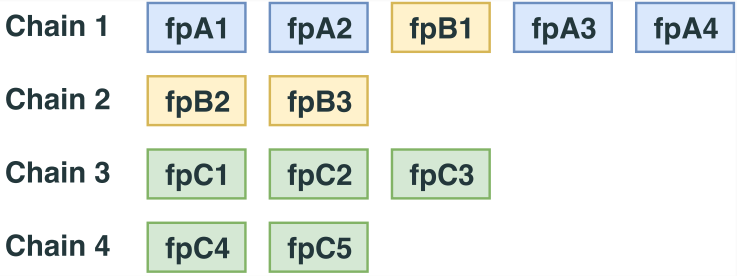
Period of time a linking algorithm correctly matches the fingerprints of a given browser in a single tracking chain



Definition of ownership

Ratio of a chain owned by the majoritarian browser

Example: $\textit{ownership}(\textit{Chain } 1) = \frac{4}{4+1} = 0.8$



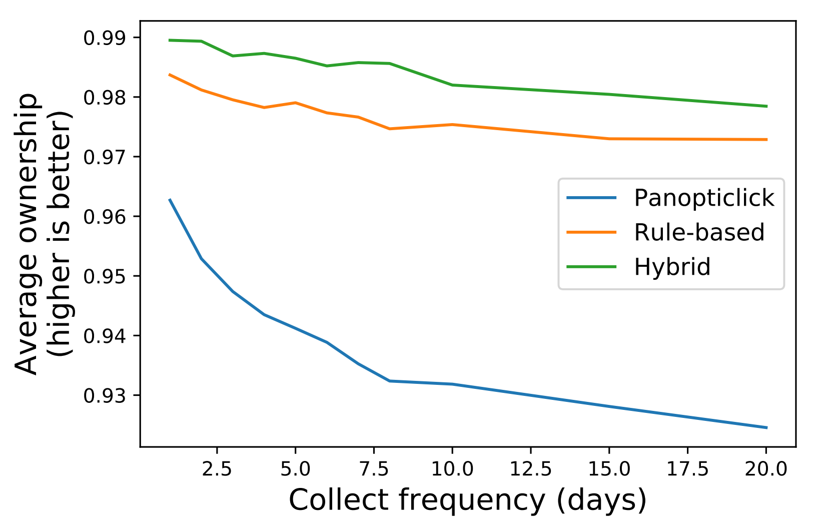
Chain 1 fpA1 fpA2 fpB1 fpA3 fpA4

Chain 2 fpB2 fpB3

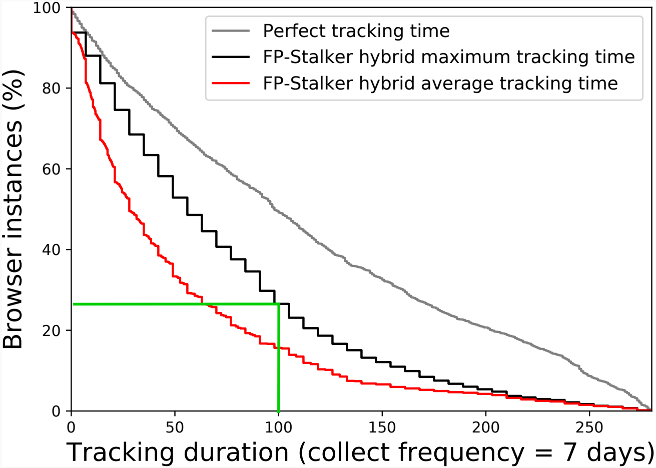
Chain 3 fpC1 fpC2 fpC3

Chain 4 fpC4 fpC5

Average ownership



Details for collect frequency = 7 days



26 % of browsers tracked more than 100 days

Conclusion

Fingerprint tracking requires uniqueness and stability

Stability depends on:

- the attributes
- the users/browsers/context

FP-Stalker, two approaches:

1. Rule-based: faster ($\approx$ 100 ms)
2. Hybrid: track 10 days longer, on average ($\approx$ 500 ms)

26% of browsers tracked more than 100 days