FP-Stalker:
Tracking Browser Fingerprint Evolutions

Antoine Vastel, Pierre Laperdrix, Walter Rudametkin, Romain Rouvoy
Wednesday 23rd May, 2018
University of Lille / Inria, France
**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

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

<table>
<thead>
<tr>
<th>Attribute</th>
<th>50th</th>
<th>90th</th>
<th>95th</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resolution</td>
<td>Never</td>
<td>3.1</td>
<td>1.8</td>
</tr>
<tr>
<td>User agent</td>
<td>39.7</td>
<td>13.0</td>
<td>8.4</td>
</tr>
<tr>
<td>Canvas</td>
<td>290.0</td>
<td>35.3</td>
<td>17.2</td>
</tr>
<tr>
<td>Language</td>
<td>Never</td>
<td>215.1</td>
<td>56.7</td>
</tr>
<tr>
<td>Accept</td>
<td>Never</td>
<td>163.8</td>
<td>109.5</td>
</tr>
<tr>
<td>Cookies</td>
<td>Never</td>
<td>Never</td>
<td>Never</td>
</tr>
</tbody>
</table>
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 $\Rightarrow$ must be *identical*
- Similarity of User agent, ..., headers $\Rightarrow$ must be $> 0.75$
- Resolution, timezone can be different
- No more than 2 attribute changes
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

<table>
<thead>
<tr>
<th>Attribute</th>
<th>FP new</th>
<th>FP database</th>
<th>Vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encoding</td>
<td>&quot;gzip, deflate, br&quot;</td>
<td>&quot;gzip, deflate&quot;</td>
<td>0.87</td>
</tr>
<tr>
<td>Languages</td>
<td>&quot;en-US, en; q=0.5&quot;</td>
<td>&quot;fr-FR, fr; q=0.8, en-US; q=0.6, en; q=0.4&quot;</td>
<td>0.53</td>
</tr>
<tr>
<td>Canvas</td>
<td><img src="image1" alt="Glyphs" />, <img src="image2" alt="Quiz" /></td>
<td><img src="image3" alt="Glyphs" />, <img src="image4" alt="Quiz" /></td>
<td>0</td>
</tr>
<tr>
<td>Number changes</td>
<td></td>
<td></td>
<td>4</td>
</tr>
</tbody>
</table>
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
Evaluate the effectiveness of browser fingerprint tracking

Test set: 59,159 fingerprints from 1,395 browsers
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: \[ \text{ownership}(\text{Chain 1}) = \frac{4}{4+1} = 0.8 \]
Average ownership

<table>
<thead>
<tr>
<th>Collect frequency (days)</th>
<th>Average ownership</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5</td>
<td>0.99</td>
</tr>
<tr>
<td>5.0</td>
<td>0.98</td>
</tr>
<tr>
<td>7.5</td>
<td>0.97</td>
</tr>
<tr>
<td>10.0</td>
<td>0.96</td>
</tr>
<tr>
<td>12.5</td>
<td>0.95</td>
</tr>
<tr>
<td>15.0</td>
<td>0.94</td>
</tr>
<tr>
<td>17.5</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Panopticlick
Rule-based
Hybrid
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