A Public and Reproducible Assessment of the Topics API on Real Data

YOHAN BEUGIN & PATRICK McDANIEL
SecWeb - May 23, 2024
Privacy Focus

Browser Privacy
Enhanced Tracking Protection
Trackers follow you around online to collect information about your browsing habits and interests. Firefox blocks many of these trackers and other malicious scripts. Learn more

Standard
Balanced for protection and performance. Pages will load normally.
Firefox blocks the following:
- Social media trackers
- Cross-site cookies in all windows
- Tracking content in Private Windows
- Cryptominers
- Fingerprinters

Includes Total Cookie Protection, our most powerful privacy feature ever
Total Cookie Protection contains cookies to the site you're on, so trackers can't use them to follow you between sites. Learn more

WebKit

Full Third-Party Cookie Blocking and More
Mar 24, 2020 by John Wilander @johnwilander

PRIVACY UPDATES
Ephemeral third-party site storage
By the Brave Privacy Team

Total Cookie Protection
Before TCP
After TCP
Privacy Sandbox
Creating a more private internet

https://privacysandbox.com
# Topics API - Overview

<table>
<thead>
<tr>
<th>User's browser</th>
<th>User's browser</th>
<th>Site that displays ads</th>
<th>Adtech code</th>
<th>Adtech code</th>
<th>Adtech code</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="User visits websites" /></td>
<td><img src="image2" alt="Browser infers topics of interest" /></td>
<td><img src="image3" alt="User visits site displaying ads" /></td>
<td><img src="image4" alt="Topics are retrieved" /></td>
<td><img src="image5" alt="Ad is requested" /></td>
<td><img src="image6" alt="Ad is displayed" /></td>
</tr>
</tbody>
</table>

**User visits websites**
The user visits websites about a range of topics, for example: "Country Music", "Makeup & Cosmetics", "Vegetarian Cuisine".

**Browser infers topics of interest**
The browser calculates the most frequently visited topics from the user’s recent browsing history.

**User visits site displaying ads**
The user visits a site whose adtech platform needs to select an ad for them.

**Topics are retrieved**
The adtech platform gets topics of interest to the user by calling the Topics API function `browsingTopics()`.

**Ad is requested**
The adtech platform uses the topics provided by the Topics API as part of the input to help select an ad.

**Ad is displayed**
An ad is displayed to the user.
Topics API - Details

Topics calculation at end of epoch $e_0$

Browsing history for $e_0$ → Topics taxonomy → Top $T=5$ topics

Call to `<browsingTopics()>` during $e_0$

API caller + Website → Random topic drawn from taxonomy $5\%$

User's top $T=5$ topics $95\%$

Topics output to API caller in random order $[t_{-2}, t_{-1}, t_{-3}]$

Origin | Topic(s)
--- | ---
www.ieee-security.org | /Computers & Electronics /News
secweb.work | /Internet & Telecom/Web Services/Web Design & Development
privacysandbox.com | /People & Society
Google’s Goals & Analyses

Privacy
1. “It must be difficult to reidentify significant numbers of users across sites using just the API.”
2. “The topics revealed by the API should be less personally sensitive about a user than what could be derived using today’s tracking methods.”

Utility
3. “The API should provide a subset of the capabilities of third-party cookies.”

Usability
4. “Users should be able to understand the API, recognize what is being communicated about them, and have clear controls. This is largely a UX responsibility but it does require that the API be designed in a way such that the UX is feasible.”
Findings

- Noisy and genuine topics can be identified
- Topics can be used to fingerprint users
- Some utility retained, but classification can be manipulated
Google's Reply

“All of the papers are using different data sets with different modeling assumptions on evolution of user interests, number of users present etc. [Google’s] research utilized real user data, while the others understandably had to generate synthetic web traces and interests […].” jkarlin

SecWeb’24 Paper

- Real browsing histories for 2,148 German users over 5 weeks (October 2018)
- New Topics API version (taxonomy, static mapping, model, etc.)
Topics Classifier

https://github.com/yohhaan/topics_classifier

This repository reproduces Google's implementations of the Topics API for the Web and for Android. This is mainly used in my research to study the privacy and utility guarantees of these proposals: PETS14 and SecWeb15.
### Real Topics Profiles

#### Initial dataset
- 2,148 users
- 9,151,243 URLs
- 49,918 unique eTLDs+1
- 67,300 unique origins

#### After filtering
- 1,207 users
- 7,746,193 URLs
- 43,684 unique eTLDs+1
- 58,370 unique origins

### Uniqueness

<table>
<thead>
<tr>
<th>Weeks</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique topics (469 topics)</td>
<td>218</td>
<td>215</td>
<td>220</td>
<td>223</td>
<td>226</td>
</tr>
<tr>
<td>Unique profiles (1,207 users)</td>
<td>1,132</td>
<td>1,132</td>
<td>1,144</td>
<td>1,145</td>
<td>1,151</td>
</tr>
</tbody>
</table>
## Real Topics Profiles

<table>
<thead>
<tr>
<th>Stability</th>
<th>Exactly 0</th>
<th>Exactly 1</th>
<th>Exactly 2</th>
<th>Exactly 3</th>
<th>Exactly 4</th>
<th>Exactly 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>From week 1 to 2</td>
<td>57 (4.7%)</td>
<td>184 (15.2%)</td>
<td>301 (24.9%)</td>
<td>373 (30.9%)</td>
<td>229 (19.0%)</td>
<td>63 (5.2%)</td>
</tr>
<tr>
<td>From week 2 to 3</td>
<td>67 (5.6%)</td>
<td>193 (16.0%)</td>
<td>315 (26.1%)</td>
<td>353 (29.2%)</td>
<td>220 (18.2%)</td>
<td>59 (4.9%)</td>
</tr>
<tr>
<td>From week 3 to 4</td>
<td>70 (5.8%)</td>
<td>188 (15.6%)</td>
<td>318 (26.3%)</td>
<td>333 (27.6%)</td>
<td>238 (19.7%)</td>
<td>60 (5.0%)</td>
</tr>
<tr>
<td>From week 4 to 5</td>
<td>70 (5.8%)</td>
<td>233 (19.3%)</td>
<td>329 (27.3%)</td>
<td>317 (26.3%)</td>
<td>215 (17.8%)</td>
<td>43 (3.6%)</td>
</tr>
</tbody>
</table>
Noise Removal - Topics Distribution on the Web

Static Mapping

CrUX 1M

Tranco 1M
# Noise Removal - Repetitions

<table>
<thead>
<tr>
<th>Epoch</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>🌟, 🎨, 🎨</td>
</tr>
<tr>
<td>1</td>
<td>🎨, 🎨, 🌟</td>
</tr>
<tr>
<td>2</td>
<td>🌟, 🌟, 🌟</td>
</tr>
<tr>
<td>3</td>
<td>🌟, 🌟, 🌟</td>
</tr>
<tr>
<td>4</td>
<td>🌟, 🌟, 🌟</td>
</tr>
<tr>
<td>5</td>
<td>🌟, 🌟, 🌟</td>
</tr>
<tr>
<td>6</td>
<td>🌟, 🌟, 🌟</td>
</tr>
</tbody>
</table>

### Simulation Results

<table>
<thead>
<tr>
<th>Week</th>
<th>Accuracy</th>
<th>Precision</th>
<th>TPR</th>
<th>FPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.954</td>
<td>0.099</td>
<td>0.793</td>
<td>0.045</td>
</tr>
<tr>
<td>4</td>
<td>0.954</td>
<td>0.100</td>
<td>0.781</td>
<td>0.045</td>
</tr>
<tr>
<td>5</td>
<td>0.954</td>
<td>0.098</td>
<td>0.744</td>
<td>0.045</td>
</tr>
</tbody>
</table>

- 🌟 noisy
- 🌟, 🎨, 🎨, 🌟, 🌟 genuine
Advertisers can Re-identify Users

Re-identification experiment

How “difficult” is it to re-identify “significant numbers of users across sites”?
Conclusion

- Topics API does not provide the same privacy to all users
- Topics API can be used to fingerprint users
- Need for reproducible evaluations
- Call for representative and anonymized topics distributions

Our Papers

- Interest-disclosing Mechanisms for Advertising are Privacy-Exposing (not Preserving) (PETS’24)
- A Public and Reproducible Assessment of the Topics API on Real Data (SecWeb’24)

Thanks!

- yohan@beugin.org
- https://yohan.beugin.org
Additional Slides
Interest-disclosing Mechanisms for Advertising are Privacy-Exposing (not Preserving) (PETS’24)
Synthetic Data Generation

Prior Web Measurement Studies:

- Replication: Why We Still Can’t Browse in Peace, Bird et al. (SOUPS’20)
- A World Wide View of Browsing the World Wide Web, Ruth et al. (IMC’22)
- Toppling top lists: evaluating the accuracy of popular website lists, Ruth et al. (IMC’22)

<table>
<thead>
<tr>
<th>Min size</th>
<th>Max size</th>
<th>N users</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25</td>
<td>21,519</td>
</tr>
<tr>
<td>26</td>
<td>50</td>
<td>11,195</td>
</tr>
<tr>
<td>51</td>
<td>75</td>
<td>6,750</td>
</tr>
<tr>
<td>76</td>
<td>100</td>
<td>4,499</td>
</tr>
<tr>
<td>101</td>
<td>125</td>
<td>2,791</td>
</tr>
<tr>
<td>126</td>
<td>150</td>
<td>1,766</td>
</tr>
<tr>
<td>151</td>
<td>-</td>
<td>3,457</td>
</tr>
</tbody>
</table>

Total 51,977

Table 1: Number of users by number of unique domain visits
Noise Removal - Topics Distribution on the Web

Static Mapping

CrUX 1M

Tranco 1M
## Noise Removal - Repetitions

<table>
<thead>
<tr>
<th>Epoch</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>★, ☮, 🌈</td>
</tr>
<tr>
<td>1</td>
<td>☮, ☮, 🌈, 📜</td>
</tr>
<tr>
<td>2</td>
<td>☮, 🌈, 🏌️️</td>
</tr>
<tr>
<td>3</td>
<td>🌈, 🏌️️, 🐦</td>
</tr>
<tr>
<td>4</td>
<td>🏌️️, ☮, 🐦</td>
</tr>
<tr>
<td>5</td>
<td>☮, 🌈, 🐦</td>
</tr>
<tr>
<td>6</td>
<td>🌈, 🏌️️, 🐦</td>
</tr>
</tbody>
</table>

### Top 5 Recovery

- **Min**
- **Median**
- **Max**

![Graph showing the size of top 5 retrieved topics over epochs](chart.png)
Advertisers can Re-identify Users

Simulation for 250k stable users

<table>
<thead>
<tr>
<th>Scenario</th>
<th>One-shot</th>
<th>Multi-shot (15-30 epochs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collusion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>25% of noisy topics removed</td>
<td>49-94% of noisy topics removed</td>
</tr>
<tr>
<td>Noise removal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Across 2 websites</td>
<td>0.4% of users re-identified 17% better than just randomly</td>
<td>57-75% of users re-identified 38-25% better than just randomly</td>
</tr>
</tbody>
</table>

Results Overview

How “difficult” is it to re-identify “significant numbers of users across sites”?
Some Utility Retained, but Topics can be Manipulated

Utility Evaluation
At least 1 true topic aligned with ground truth in about 60% of cases

Misclassification
Topics (word): Comics (batman), Dance (dance), …
Domain: example.com
Crafted Subdomains: batman.example.com, dance.example.com, …

350 topics × top 10k domains = 3.5M subdomains