Topics API for the Web

- Third-party cookies
- Fingerprinting
- Interest-based advertising

Google’s Goals

1. It must be difficult to reidentify significant numbers of users across sites using just the API.
2. The API should provide a subset of the capabilities of third-party cookies.
3. The topics revealed by the API should be less personally sensitive about a user than what could be derived using today’s tracking methods.
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.

Privacy & Utility Analysis of the Topics API for the Web

The Topics API is used to fingerprint users by returning a set of topics associated with a user based on their browsing history. The API caller (e.g., CrUX) is responsible for calling the `browsingTopics()` function during the user’s browsing history. The Topics output is then passed to the API caller in a random order.

### Topics Distribution on Top 1M Websites

- Asymmetric topics distribution on the web: Our classifier considers every topic that does not appear at least on 10 websites among the top 1M as noisy.

- Repetitions leak real topics: Coupon Collector’s Problem.
  - One-shot: 25% of noisy topics removed.
  - Multi-shot: 49-94% (15-30 epochs) removed.

Result: Plausible deniability can be refuted.

### Identification of Noisy and Real Topics

- Real topics:batman, dance, comics, etc.
- Noisy topics: batman.example.com, dance.example.com, etc.

#### Threat Model

- Attackers can manipulate topics to fingerprint users.
- The API can be designed in a way such that the UX is feasible.

#### Takeaways

- Topics can be used to fingerprint users.
- Users have stable and unique web behaviors that need to be considered.
- Google’s non-reproducible analyses are disconnected from reality and lack systematization in their approach.

### Need for a (Research) Sandstorm through the Privacy Sandbox

- Call for reproducible analyses and release of tools and datasets.
- More evaluations are required to understand all potential impacts.
- Launch of a new research hub at: https://privacysandstorm.com

#### Key Takeaways

- Topics can not guarantee non-reidentification across websites to all users.
- Users have stable and unique web behaviors that need to be considered.
- Google’s non-reproducible analyses are disconnected from reality and lack systematization in their approach.

### Systematic and Reproducible Analysis

- Goals redefined to be quantifiable
- Formal and worst-case studies
- Measurements

#### Measurement on Real Browsing Histories

- Real data: 1207 users from Germany over 5 weeks in October 2018.
- Uniqueness: 94% have unique topics profiles.
- Stability: at least 47% have 3 or more stable topics.

### Re-identification Experiment

- Simulation: quantification of the fingerprinting risk of Topics on an arbitrarily large population of users (250k) over time (30 epochs).
- k-anonymity across time: How “difficult” is it to re-identify “significant numbers of users across sites”?

Result: users can be fingerprinted by the Topics API.

### Classification Comparison

- Manual Verification (385 domains)
- Static Mapping (10k domains)
- Cloudflare Radar Categorization (348k domains)

Result: at least 1 true topic aligned with ground truth in about 60% of cases.

### Abuse Potential

- Topics (word): Comics (batman), Dance (dance), ...
- Domain: example.com, ...
- Crafted subdomains: batman.example.com, dance.example.com, ...

350 topics x top 10k domains = 3.5M subdomains