Interest-disclosing Mechanisms for Advertising are Privacy-Exposing

(not Preserving)

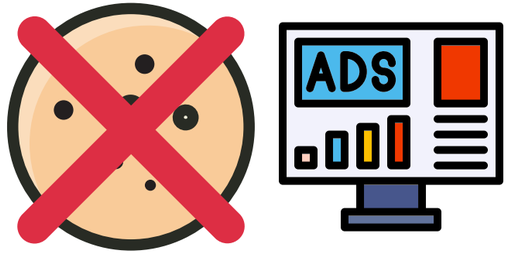
Yohan Beugin & Patrick McDaniel

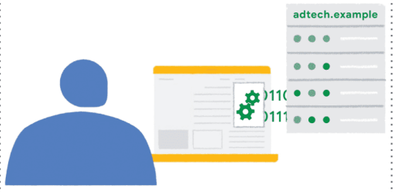
PETS - July 18, 2024

Computer Sciences
SCHOOL OF COMPUTER, DATA & INFORMATION SCIENCES
UNIVERSITY OF WISCONSIN–MADISON

MADS&P

Topics API - Overview

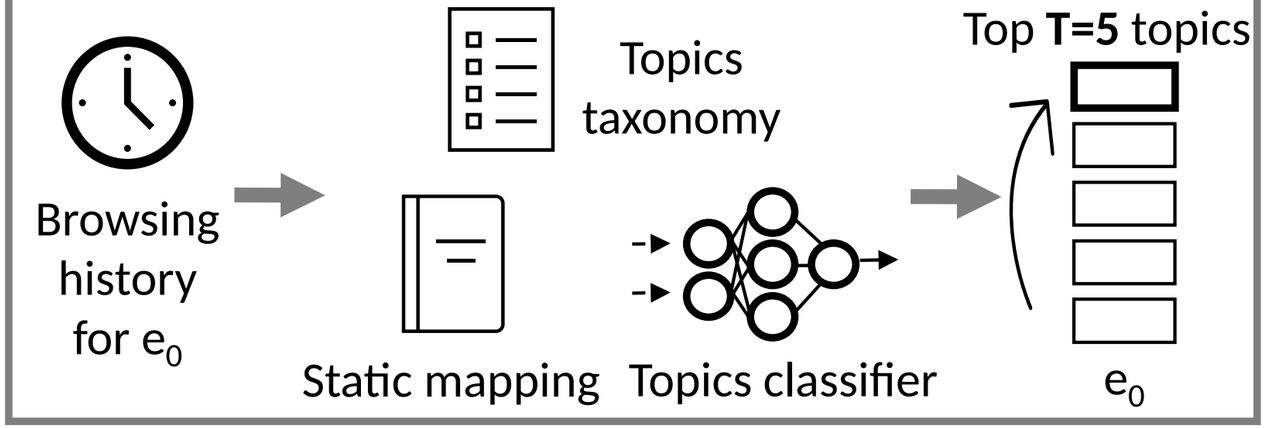


User's browser	User's browser	Site that displays ads	Adtech code	Adtech code	Adtech code
					
User visits websites	Browser infers topics of interest	User visits site displaying ads	Topics are retrieved	Ad is requested	Ad is displayed
The user visits websites about a range of topics, for example: "Country Music", "Makeup & Cosmetics", "Vegetarian Cuisine"	The browser calculates the most frequently visited topics from the user's recent browsing history	The user visits a site whose adtech platform needs to select an ad for them	The adtech platform gets topics of interest to the user by calling the Topics API function `browsingTopics()`	The adtech platform uses the topics provided by the Topics API as part of the input to help select an ad	An ad is displayed to the user

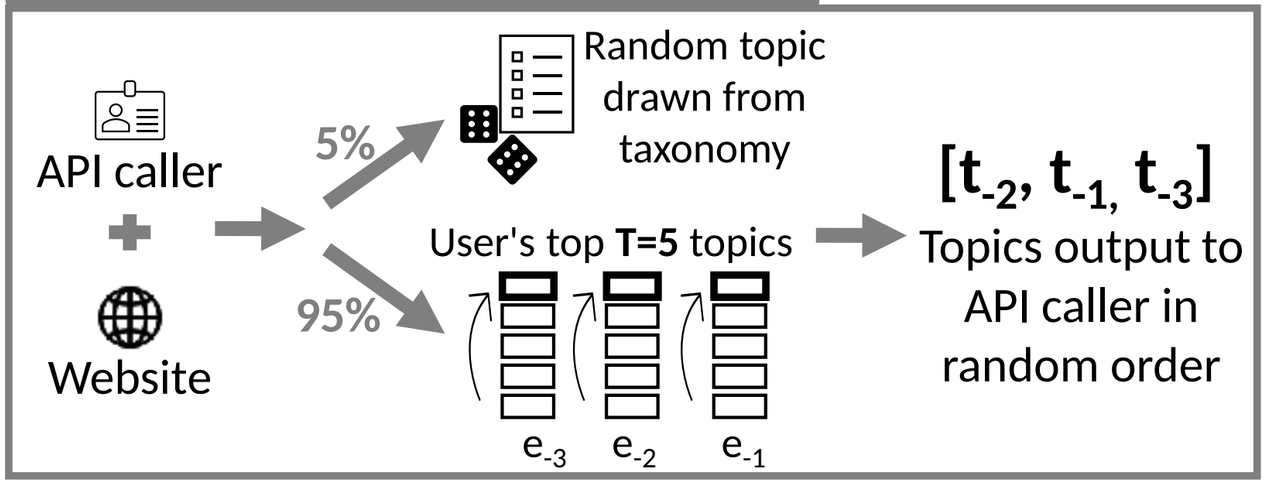
Overview

Topics API - Details

Topics calculation at end of epoch e_0



Call to <browsingTopics()> during e_0



Origin	Topic(s)
petsymposium.org	/Pets & Animals/Pets /Pets & Animals
privacysandbox.com	/People & Society
www.bristol.ac.uk	/Jobs & Education/Education/Colleges & Universities

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."*

Lack of a systematic and reproducible evaluation quantifying these goals

How to evaluate the Topics API?

1. Goals redefined to be quantifiable

2. Statistical analysis and observations

3. Measurements

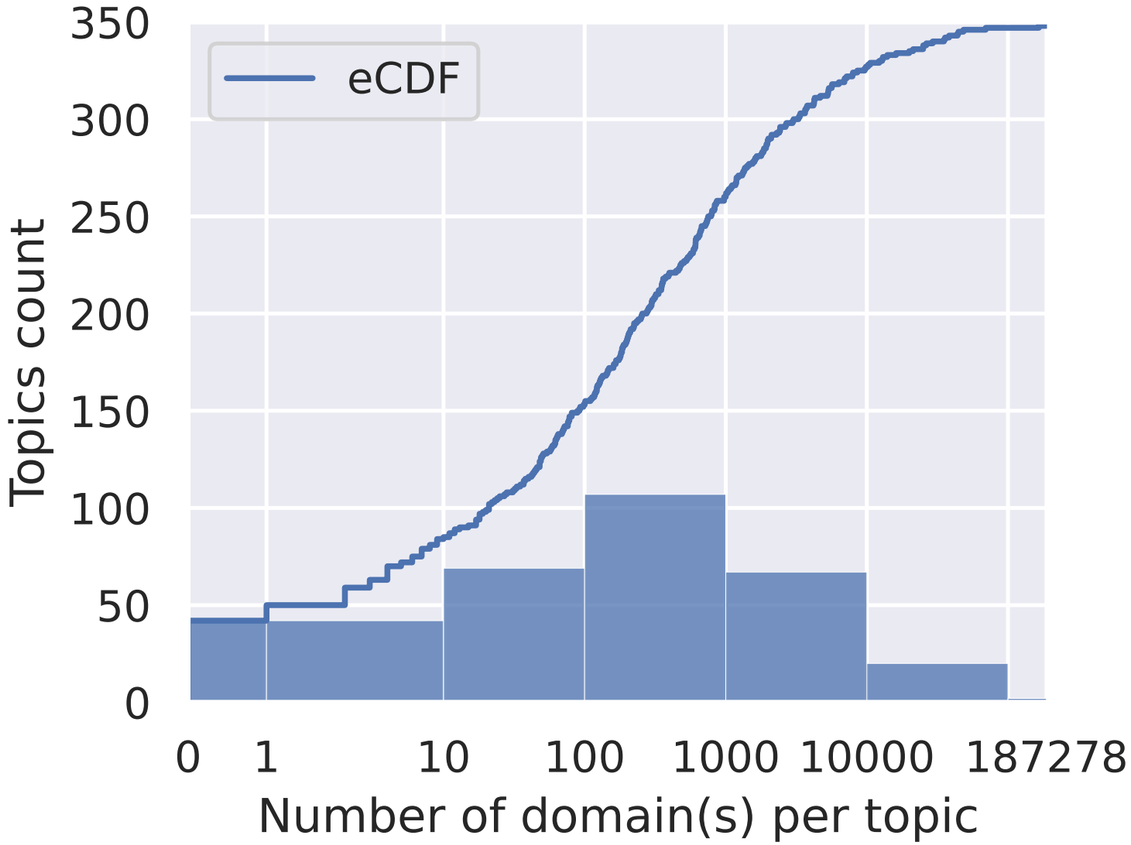
4. Empirical analysis (worst-case)

Noise Identification & Plausible Deniability Refutation

Repetitions Leak Genuine Topics

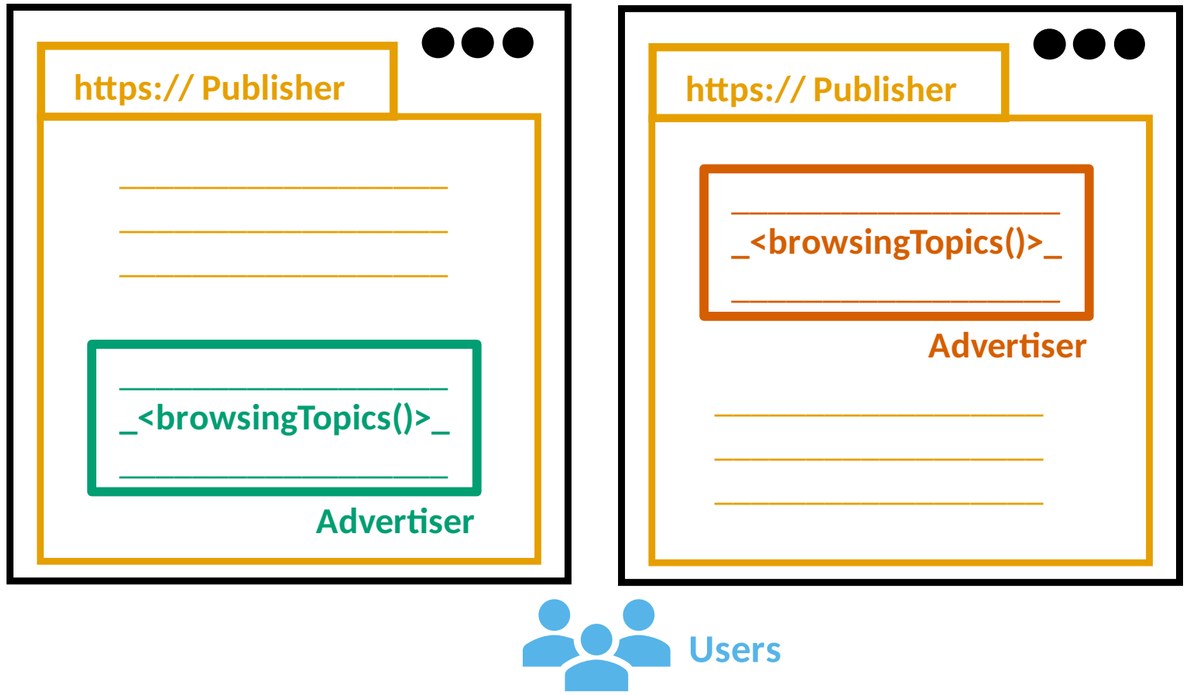
Epoch	Topics
0	
1	
2	
3	
4	
5	
6	

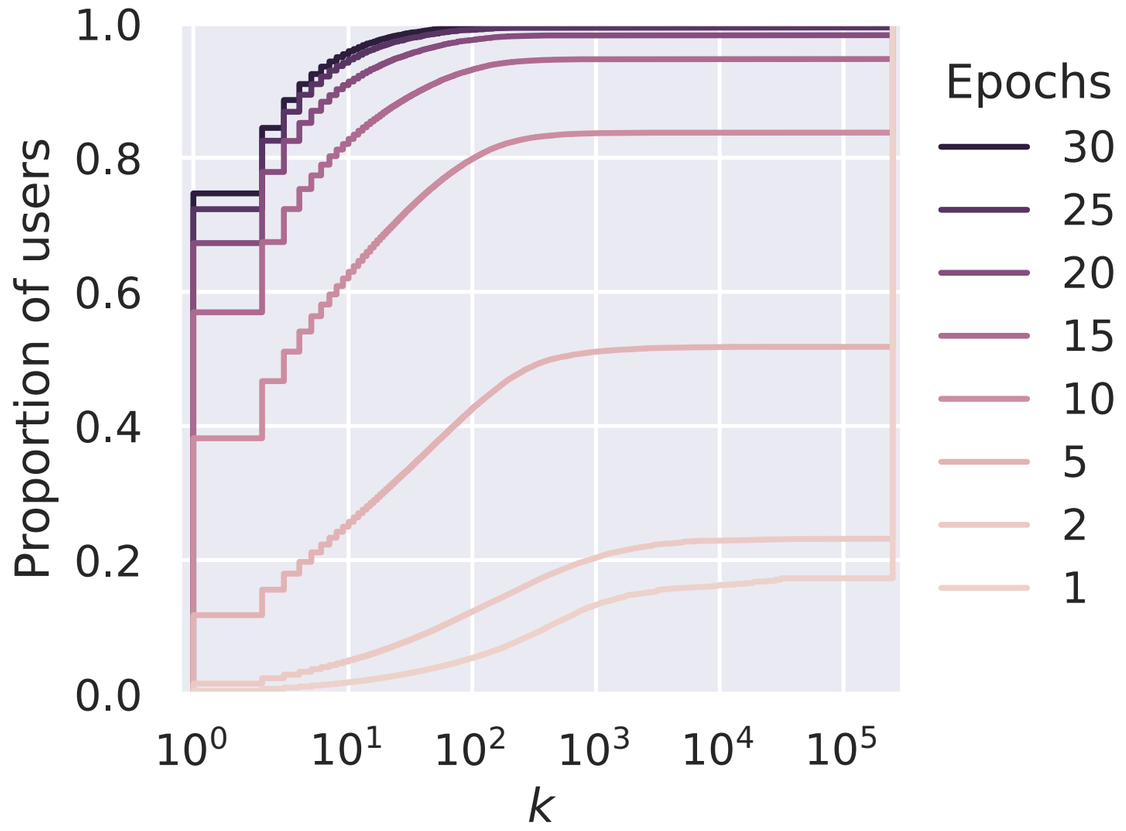
Asymmetric Topics Distribution



CrUX 1M

Users can be Tracked across Websites





250k users simulation

How *"difficult"* is it to re-identify *"significant numbers of users across sites"*?

Some Utility Retained, but Topics can be Manipulated

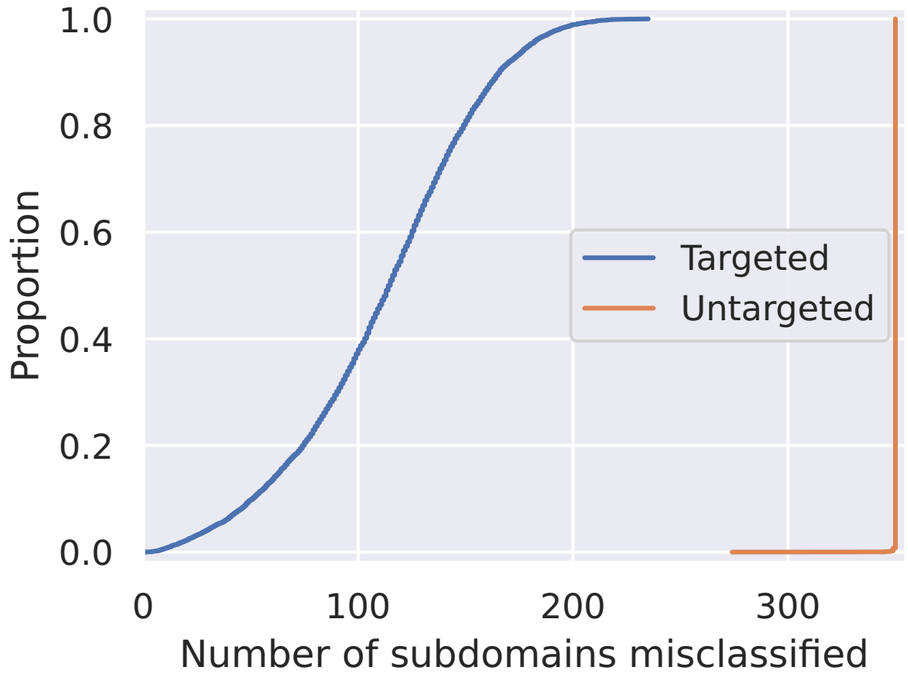
Comparison Result

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



Our Follow-up Analysis (SecWeb'24)

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

Yohan Beugin
University of Wisconsin-Madison
Madison, USA
ybeugin@cs.wisc.edu

Patrick McDaniel
University of Wisconsin-Madison
Madison, USA
mcdaniel@cs.wisc.edu

***Abstract*—The `Topics` API for the web is Google's privacy-enhancing alternative to replace third-party cookies. Results of prior work have led to an ongoing discussion between Google and research communities about the capability of `Topics` to trade off both utility and privacy. The central point of contention is largely around the realism of the datasets used in these analyses and their reproducibility; researchers using data collected on a small sample of users or generating synthetic datasets, while Google's results are inferred from a private dataset. In this paper, we complement prior research by performing a reproducible assessment of the latest version of the `Topics` API on the largest and publicly available dataset of real browsing histories. First, we measure how unique and stable real users' interests are over time. Then, we evaluate if `Topics` can be used to fingerprint the users from these real browsing traces by adapting methodologies from prior privacy studies. Finally, we call on web actors to perform and enable reproducible evaluations by releasing anonymized distributions. We find that 46 %, 55 %, and 60 % of the 1207 users in the dataset are uniquely re-identified across websites after only 1, 2, and 3 observations of their topics by advertisers, respectively. This paper shows on real data that `Topics` does not provide the same privacy guarantees to all users, further highlighting the need for public and reproducible evaluations of the claims made by new web proposals.**

bility of the `Topics` API to deliver on both its utility and privacy objectives. The major point of contention between the analyses carried out by Google and researchers is the access asymmetry to real browsing data as well as the resulting reproducibility of the evaluations. Indeed, while researchers have either collected browsing data on a small sample of 268 users [5] or synthetically generated large traces [6], Google performed their evaluations on a private dataset [7], [8] and only reported aggregate results [4], making it impossible to reproduce their evaluation.

In this paper, we evaluate the latest version of the `Topics` API on the largest publicly available dataset of real browsing histories that we could find [9], [10]; complementing prior work and proposing an alternative to having to trust Google's non-reproducible assertions. We adapt prior methodologies to measure the fingerprinting potential of the `Topics` API on this publicly accessible dataset. Finally, we discuss future research avenues and call on web actors to release anonymized distributions to enable further reproducible analyses.

First, we measure on an anonymized dataset of over a month of real browsing histories how stable and unique users' online behaviors and interests are over time. This is to compare with a stability assumption assumed in prior work [6]. Then, we adapt prior privacy analyses of the `Topics` API, but on the latest version of the proposal[1] as a new topics taxonomy, a new machine learning classifier,

Findings on Real Users

- 1207 German users over 5 weeks (2018)
- Stable and unique topics
- New Topics API version

Topic observation(s)	Re-identified
1	46%
2	55%
3	60%

Conclusion

- The Topics API can be used to fingerprint users
- Some utility is retained, but classification can be manipulated
- Need for reproducible evaluations (topics distribution, testbed, …)

Publications & Artifacts

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

Thanks!

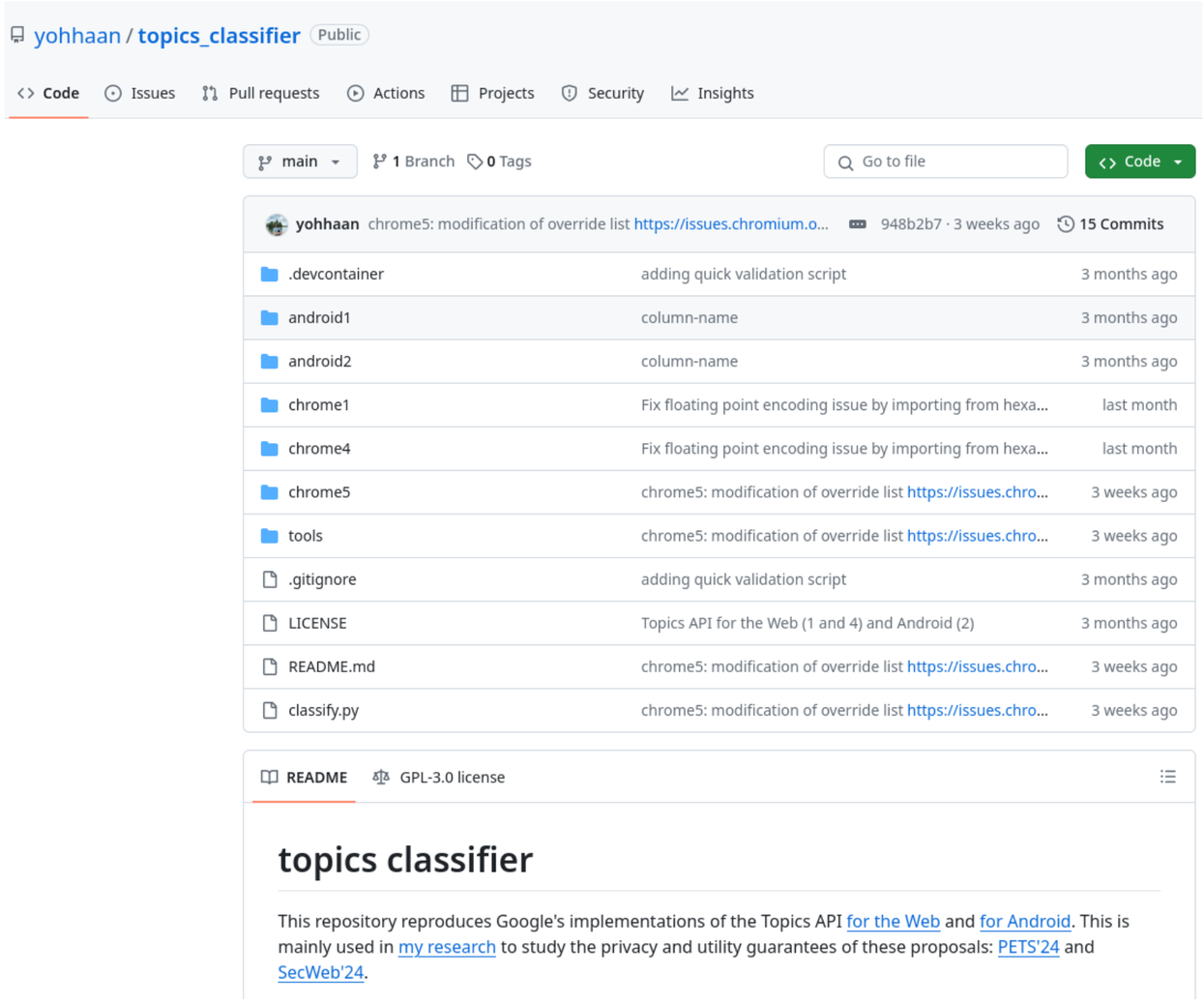
HotPETS'24 (Tomorrow!): The Need for a (Research) Sandstorm through the Privacy Sandbox

yohan@beugin.org https://yohan.beugin.org

Additional Slides

Interest-disclosing Mechanisms for Advertising are Privacy-Exposing (not Preserving)   (PETS'24)

Classifier



https://github.com/yohhaan/topics_classifier

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)

Min size	Max size	N users
1	25	21,519
26	50	11,195
51	75	6,750
76	100	4,499
101	125	2,791
126	150	1,766
151	-	3,457
	Total	51,977

Table 1: Number of users by number of unique domain visits

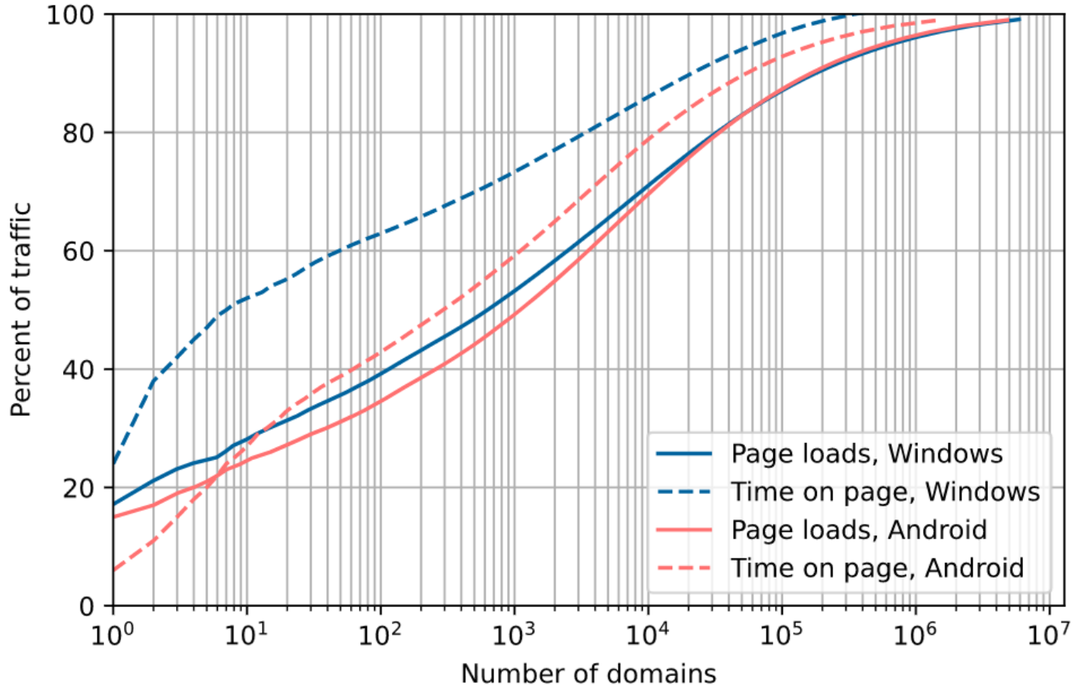
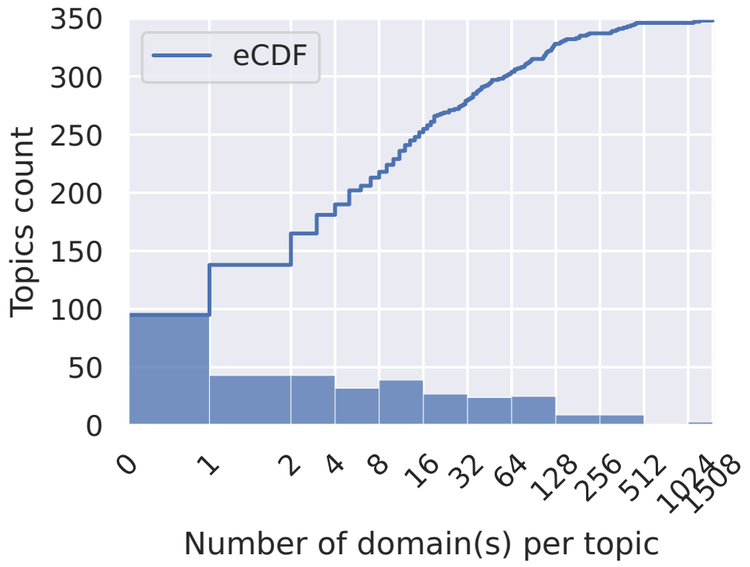
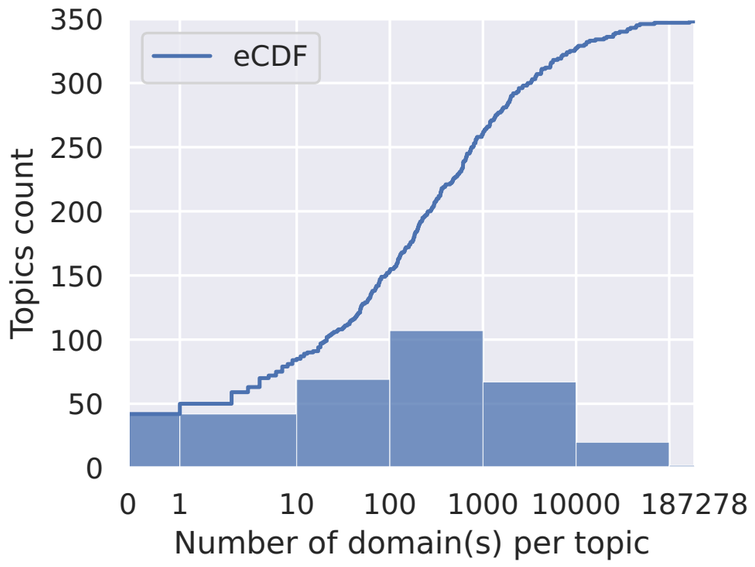


Figure 1: Distribution of Web Traffic By Website Rank

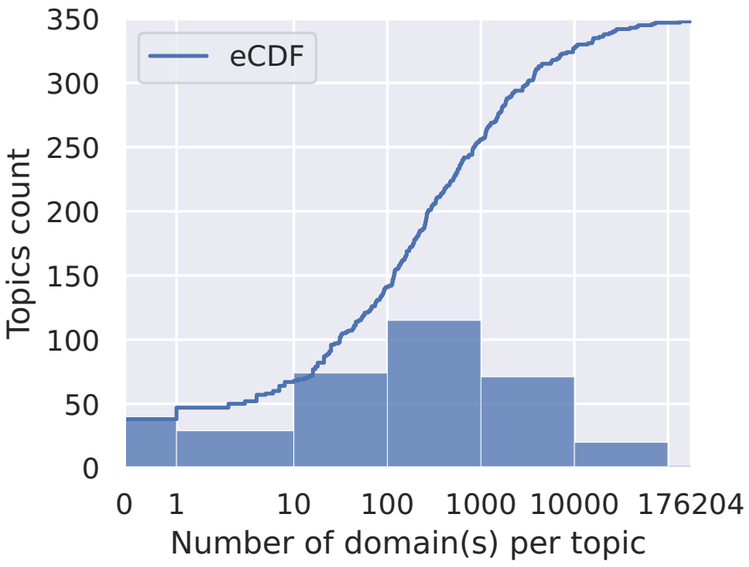
Results



Static Mapping



CrUX 1M



Tranco 1M

Collusion \ Scenario	One-shot	Multi-shot (15-30 epochs)
None **Noise removal**	25% of noisy topics removed	49-94% of noisy topics removed
Across 2 websites **Cross-site tracking**	0.4% of users re-identified 17% better than just randomly	57-75% of users re-identified 38-25% better than just randomly

A Public and Reproducible Assessment of the Topics API on Real Data   (SecWeb'24)

Ongoing Discussion & Contention

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.)

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

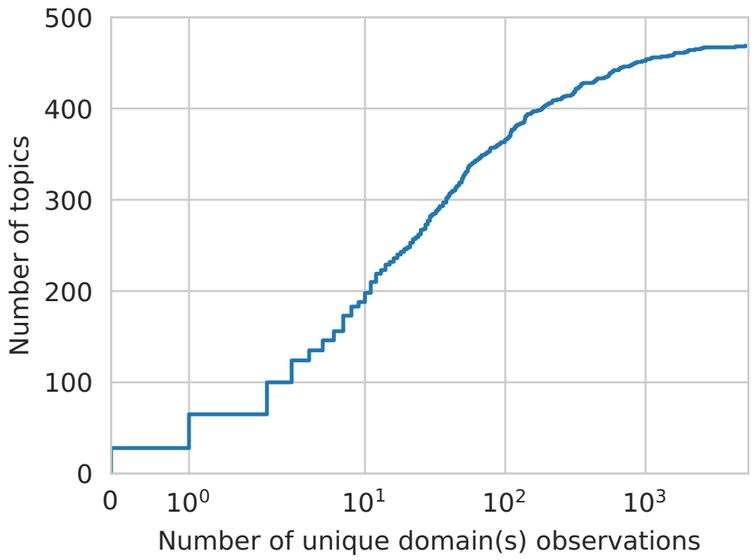
Weeks	1	2	3	4	5
Unique topics (469 topics)	218	215	220	223	226
Unique profiles (1 207 users)	1 132	1 132	1 144	1 145	1 151

Real Topics Profiles

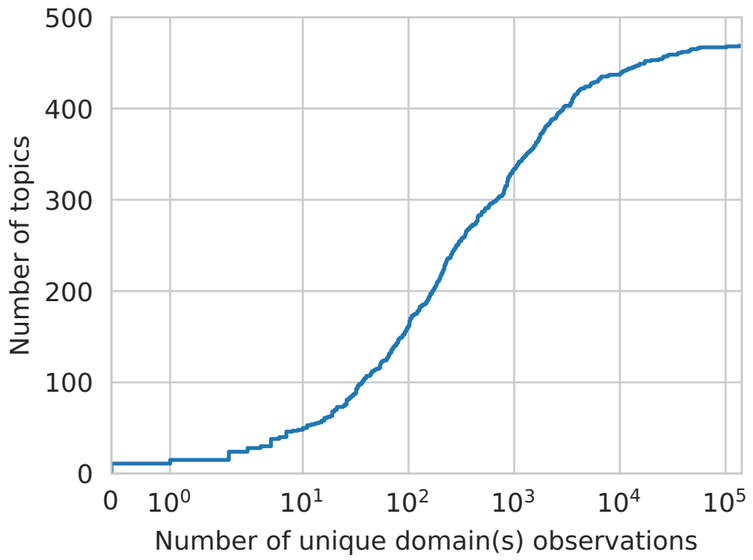
Stability

	Exactly 0	Exactly 1	Exactly 2	Exactly 3	Exactly 4	Exactly 5
From week 1 to 2	57 (4.7%)	184 (15.2%)	301 (24.9%)	373 (30.9%)	229 (19.0%)	63 (5.2%)
From week 2 to 3	67 (5.6%)	193 (16.0%)	315 (26.1%)	353 (29.2%)	220 (18.2%)	59 (4.9%)
From week 3 to 4	70 (5.8%)	188 (15.6%)	318 (26.3%)	333 (27.6%)	238 (19.7%)	60 (5.0%)
From week 4 to 5	70 (5.8%)	233 (19.3%)	329 (27.3%)	317 (26.3%)	215 (17.8%)	43 (3.6%)

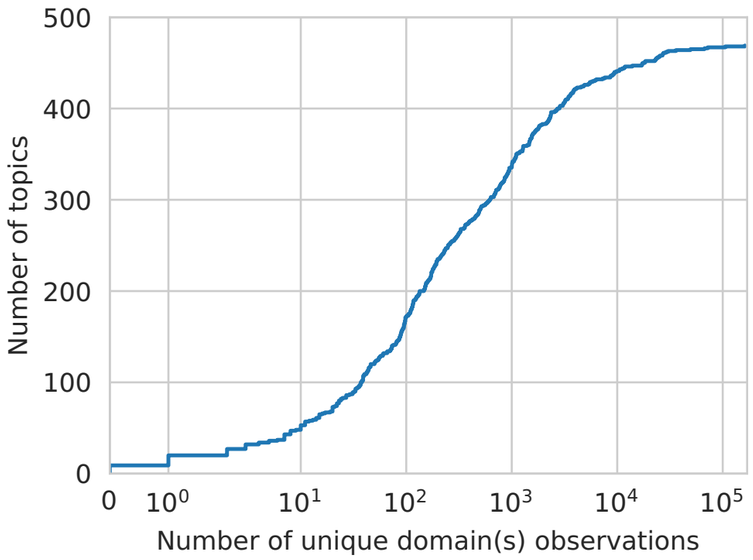
Noise Removal - Topics Distribution on the Web



Static Mapping



CrUX 1M



Tranco 1M

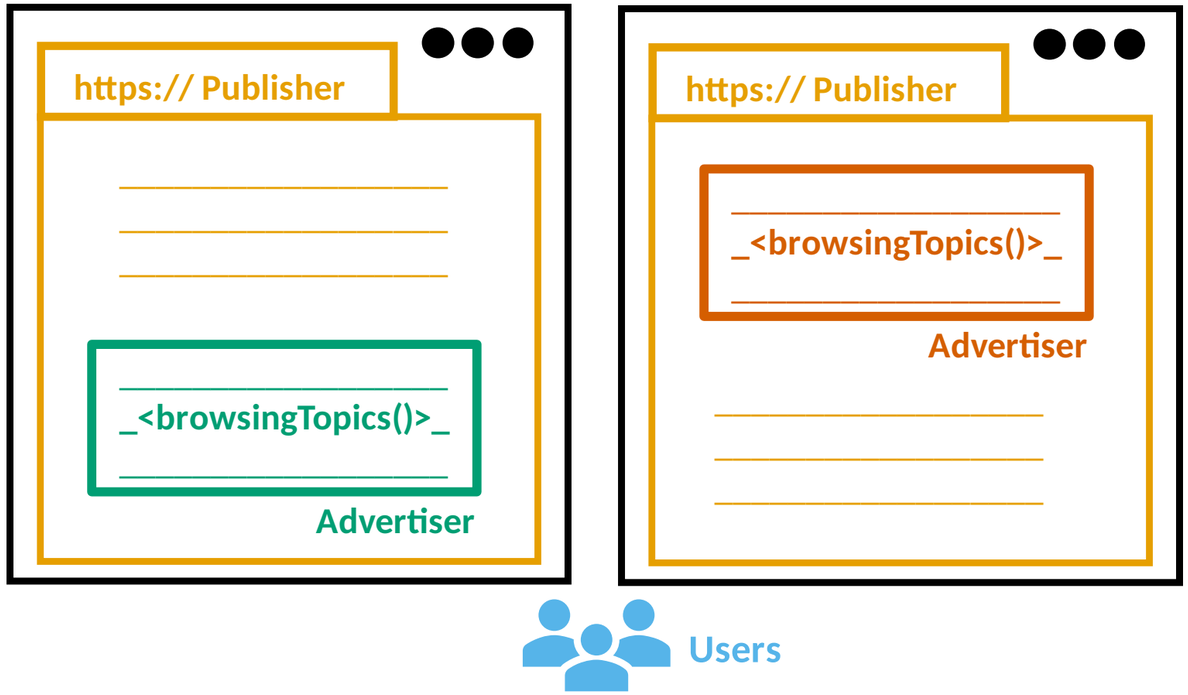
Noise Removal - Repetitions

Epoch	Topics
0	
1	
2	
3	
4	
5	
6	

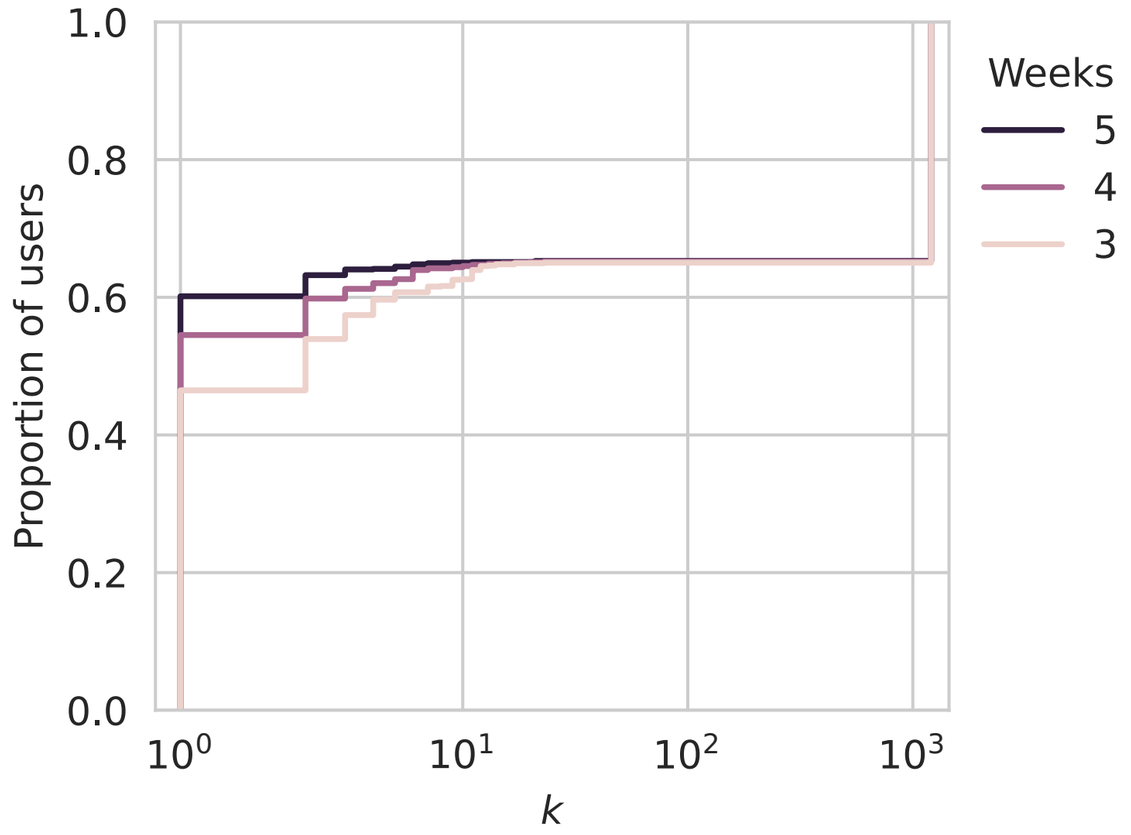
Simulation Results

Week	Accuracy	Precision	TPR	FPR
3	0.954	0.099	0.793	0.045
4	0.954	0.100	0.781	0.045
5	0.954	0.098	0.744	0.045

Advertisers can Re-identify Users



Re-identification experiment



How *"difficult"* is it to re-identify *"significant numbers of users across sites"*?