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

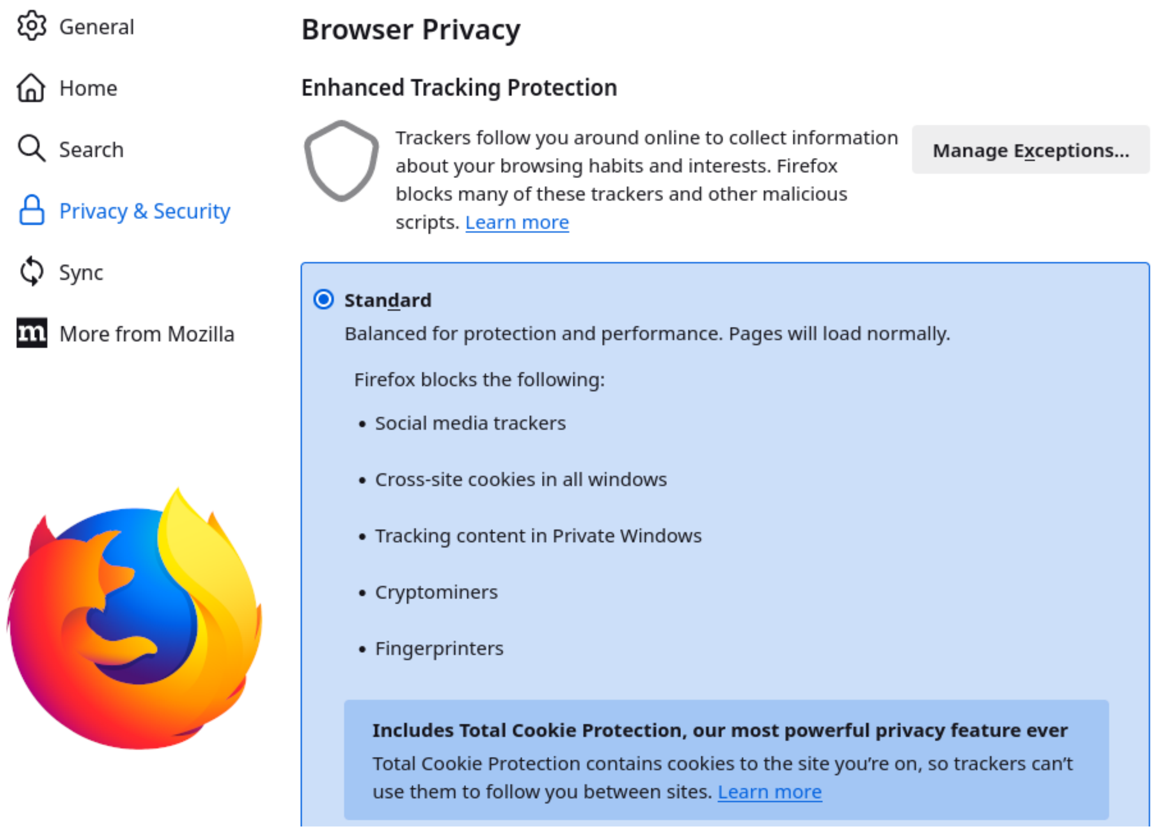
YOHAN BEUGIN & PATRICK MCDANIEL

SecWeb - May 23, 2024

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

MADS&P

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

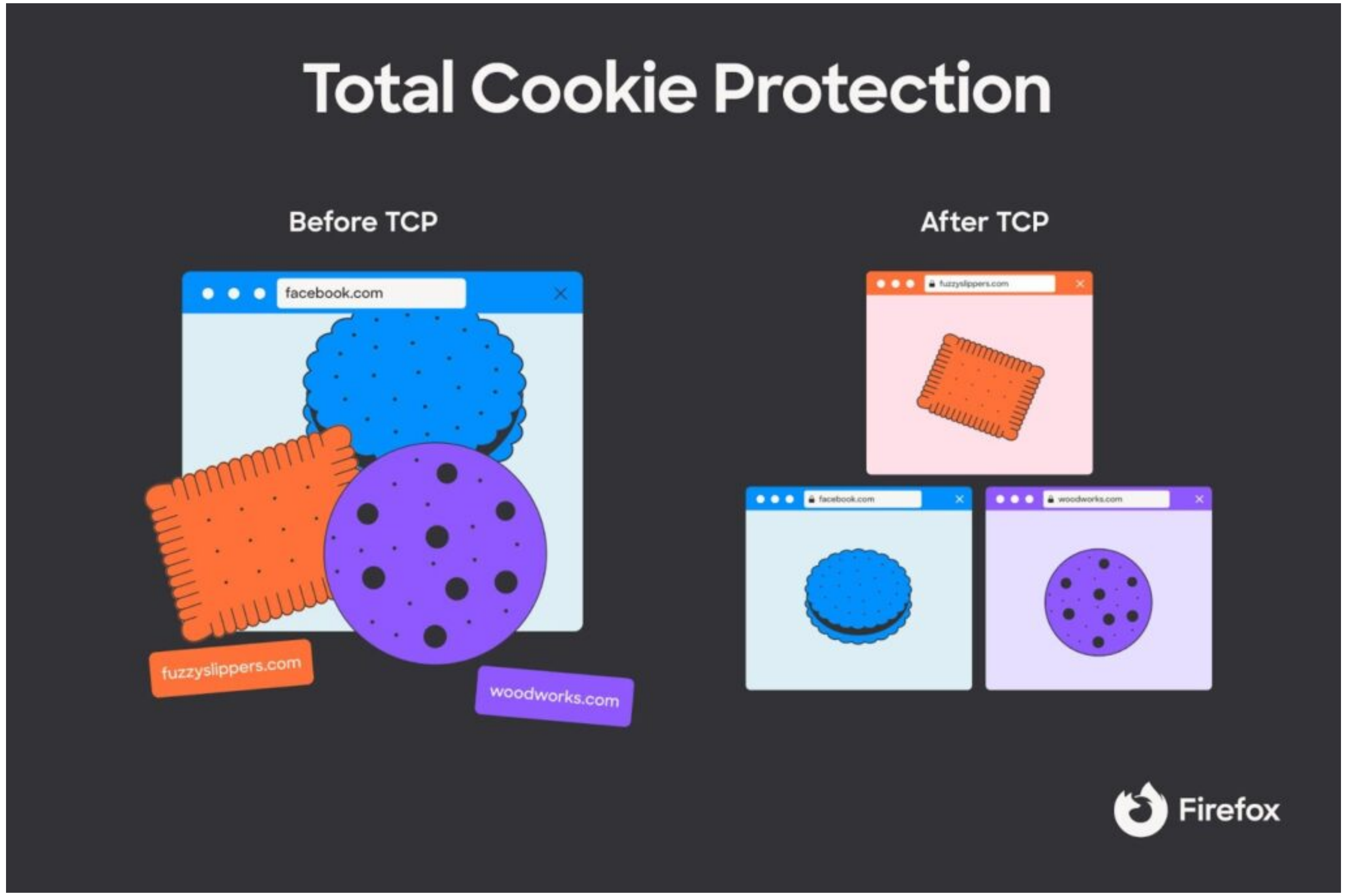
Full Third-Party Cookie Blocking and More

Mar 24, 2020 by John Wilander @johnwilander

PRIVACY UPDATES

Ephemeral third-party site storage

By the Brave Privacy Team

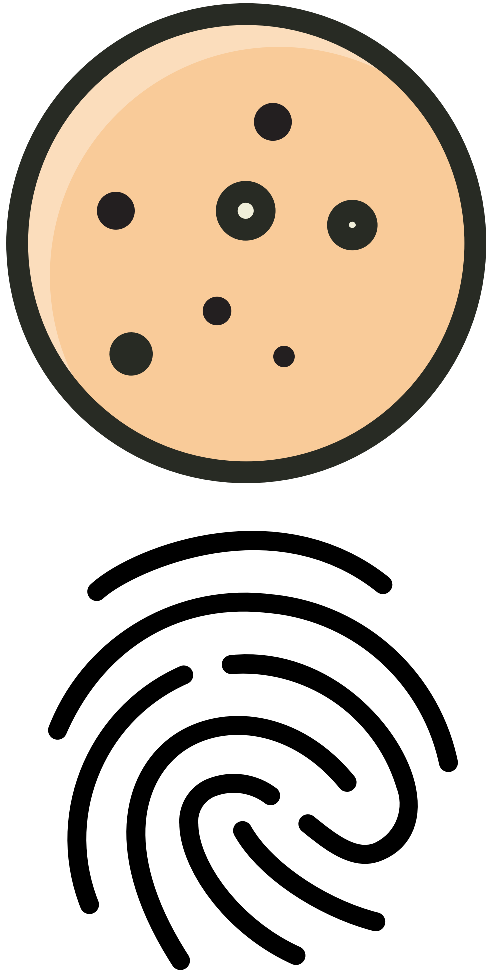


Total Cookie Protection

The Privacy Sandbox from Google

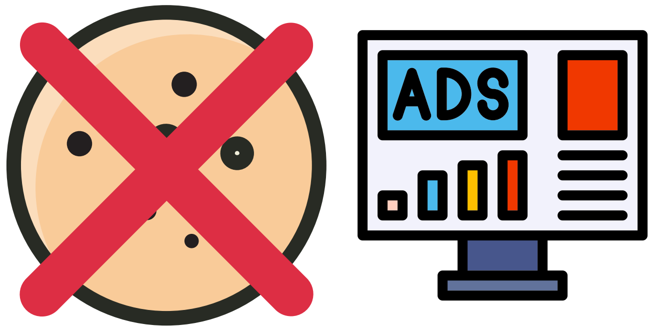
Privacy Sandbox

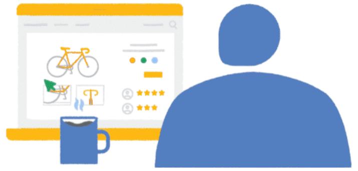
Creating a more private internet



https://privacysandbox.com

Topics API - Overview

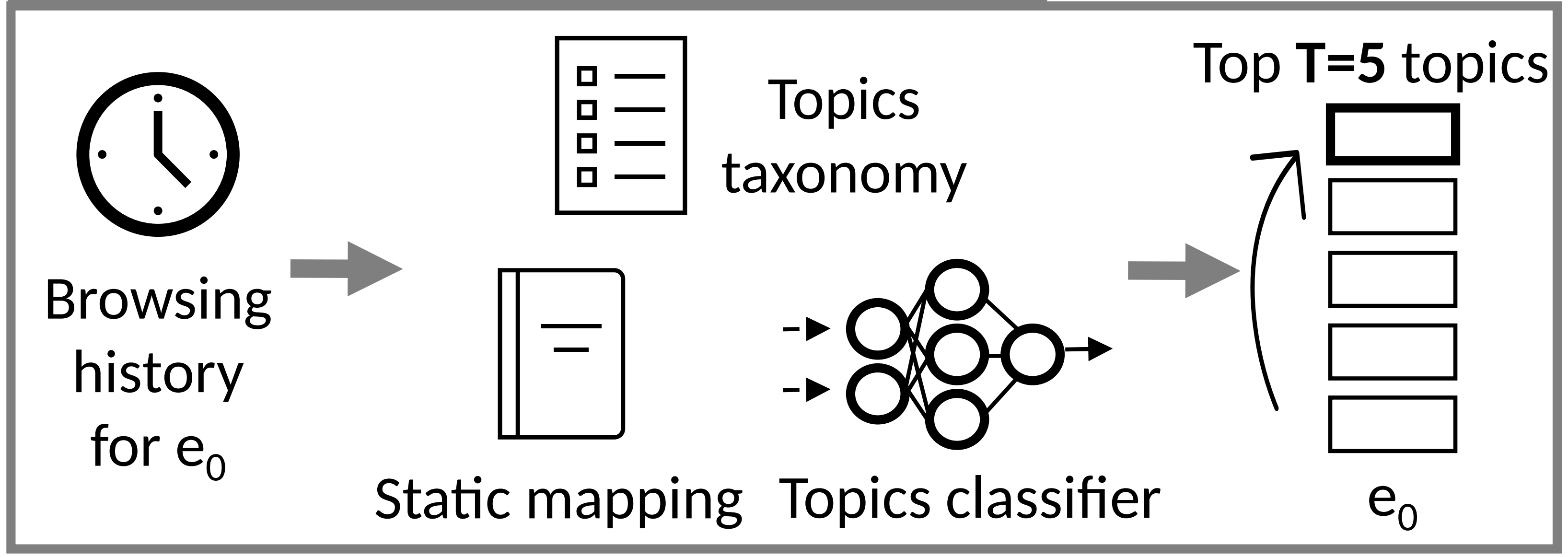


User's browser	User's browser	Site that displays ads	Adtech code	Adtech code	Adtech code
					
User visits websites →	Browser infers topics of interest →	User visits site displaying ads →	Topics are retrieved →	Ad is requested →	Ad is displayed
The user visits websites about a range of topics, for example: "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



Browsing history for e_0

Topics taxonomy

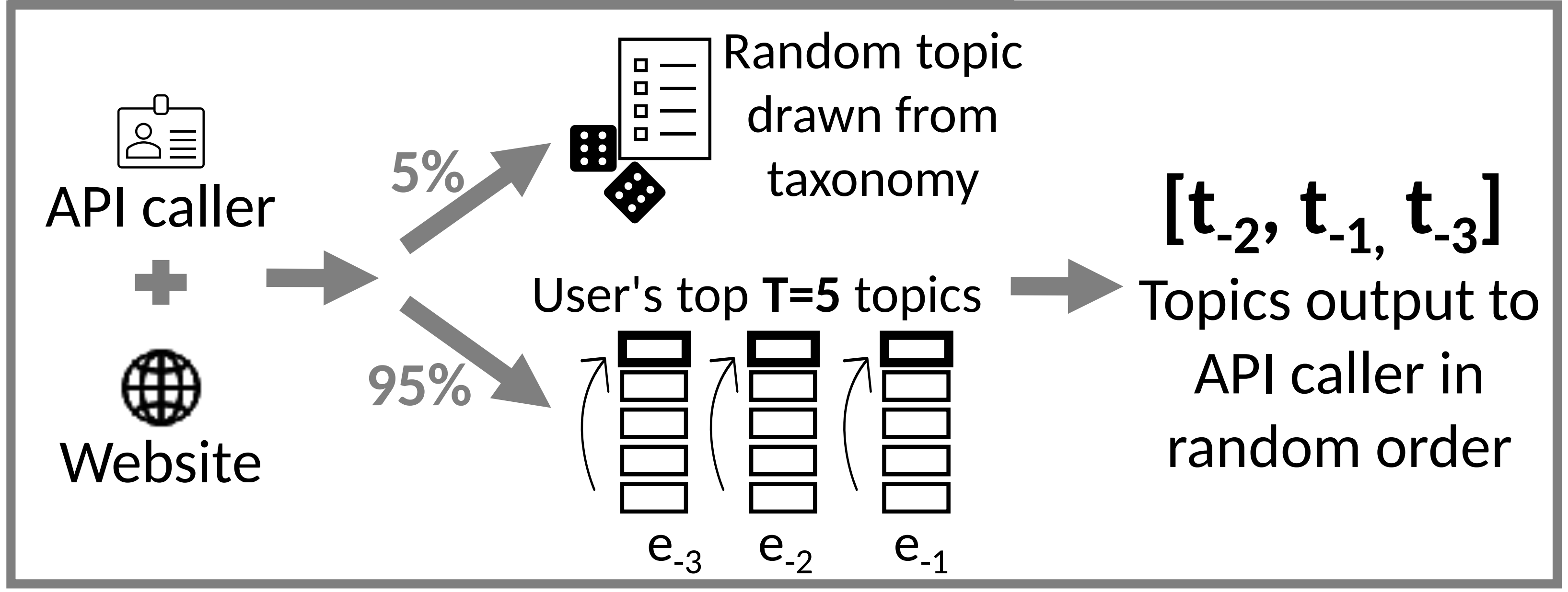
Static mapping

Topics classifier

Top **T=5** topics

e_0

Call to <browsingTopics()> during e_0



API caller

Website

5%

95%

Random topic drawn from taxonomy

User's top **T=5** topics

e_{-3} e_{-2} e_{-1}

$[t_{-2}, t_{-1}, t_{-3}]$

Topics output to API caller in random order

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."*

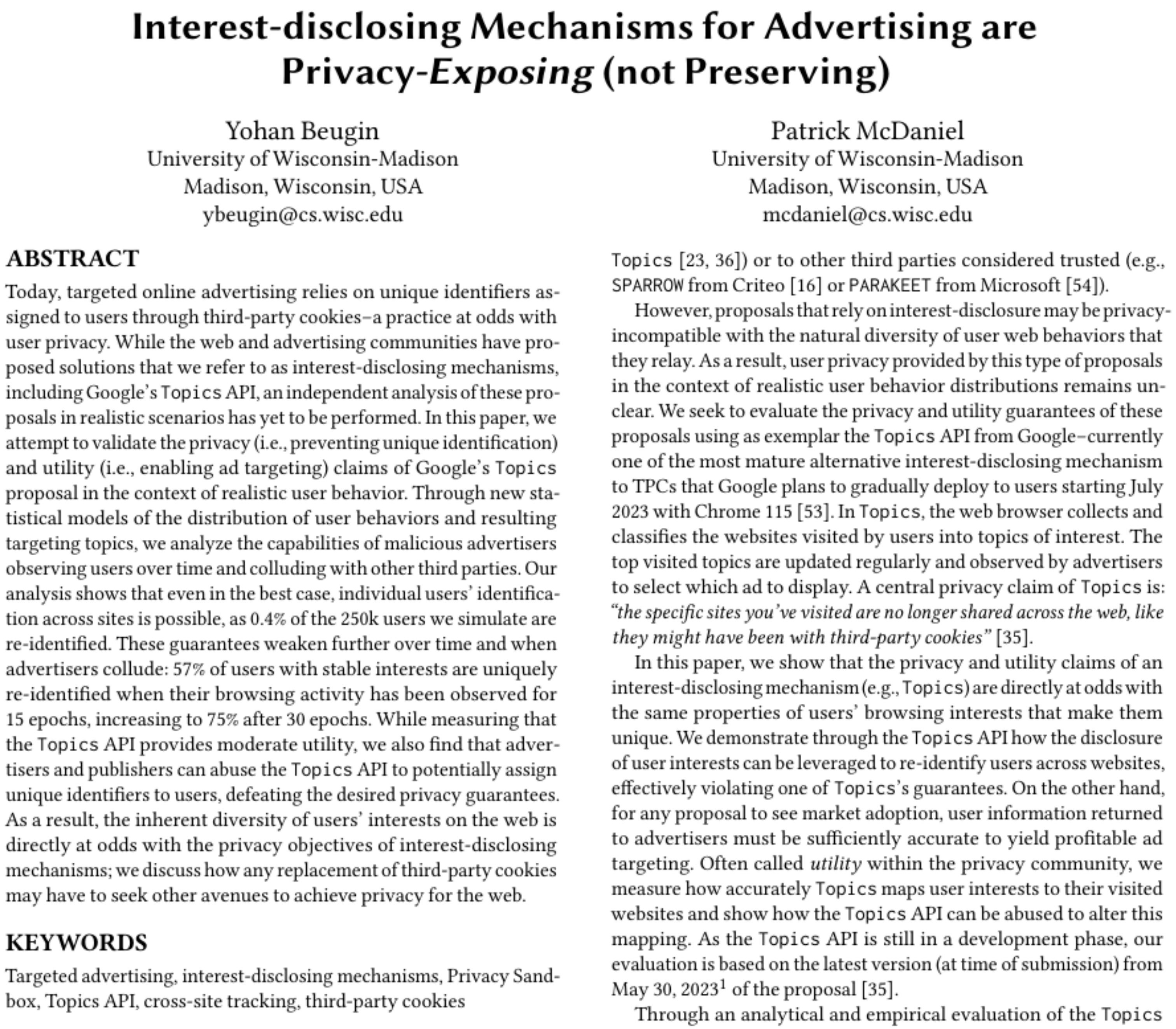
Utility

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."*

Our Prior Analysis (PETS’24)



Interest-disclosing Mechanisms for Advertising are Privacy-*Exposing* (not Preserving)

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

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

ABSTRACT

Today, targeted online advertising relies on unique identifiers assigned to users through third-party cookies–a practice at odds with user privacy. While the web and advertising communities have proposed solutions that we refer to as interest-disclosing mechanisms, including Google’s Topics API, an independent analysis of these proposals in realistic scenarios has yet to be performed. In this paper, we attempt to validate the privacy (i.e., preventing unique identification) and utility (i.e., enabling ad targeting) claims of Google’s Topics proposal in the context of realistic user behavior. Through new statistical models of the distribution of user behaviors and resulting targeting topics, we analyze the capabilities of malicious advertisers observing users over time and colluding with other third parties. Our analysis shows that even in the best case, individual users’ identification across sites is possible, as 0.4% of the 250k users we simulate are re-identified. These guarantees weaken further over time and when advertisers collude: 57% of users with stable interests are uniquely re-identified when their browsing activity has been observed for 15 epochs, increasing to 75% after 30 epochs. While measuring that the Topics API provides moderate utility, we also find that advertisers and publishers can abuse the Topics API to potentially assign unique identifiers to users, defeating the desired privacy guarantees. As a result, the inherent diversity of users’ interests on the web is directly at odds with the privacy objectives of interest-disclosing mechanisms; we discuss how any replacement of third-party cookies may have to seek other avenues to achieve privacy for the web.

KEYWORDS

Targeted advertising, interest-disclosing mechanisms, Privacy Sandbox, Topics API, cross-site tracking, third-party cookies

Topics [23, 36]) or to other third parties considered trusted (e.g., SPARROW from Criteo [16] or PARAKEET from Microsoft [54]).

However, proposals that rely on interest-disclosure may be privacy-incompatible with the natural diversity of user web behaviors that they relay. As a result, user privacy provided by this type of proposals in the context of realistic user behavior distributions remains unclear. We seek to evaluate the privacy and utility guarantees of these proposals using as exemplar the Topics API from Google–currently one of the most mature alternative interest-disclosing mechanism to TPCs that Google plans to gradually deploy to users starting July 2023 with Chrome 115 [53]. In Topics, the web browser collects and classifies the websites visited by users into topics of interest. The top visited topics are updated regularly and observed by advertisers to select which ad to display. A central privacy claim of Topics is: *“the specific sites you’ve visited are no longer shared across the web, like they might have been with third-party cookies”* [35].

In this paper, we show that the privacy and utility claims of an interest-disclosing mechanism (e.g., Topics) are directly at odds with the same properties of users’ browsing interests that make them unique. We demonstrate through the Topics API how the disclosure of user interests can be leveraged to re-identify users across websites, effectively violating one of Topics’s guarantees. On the other hand, for any proposal to see market adoption, user information returned to advertisers must be sufficiently accurate to yield profitable ad targeting. Often called *utility* within the privacy community, we measure how accurately Topics maps user interests to their visited websites and show how the Topics API can be abused to alter this mapping. As the Topics API is still in a development phase, our evaluation is based on the latest version (at time of submission) from May 30, 2023[1] of the proposal [35].

Through an analytical and empirical evaluation of the Topics

Findings

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

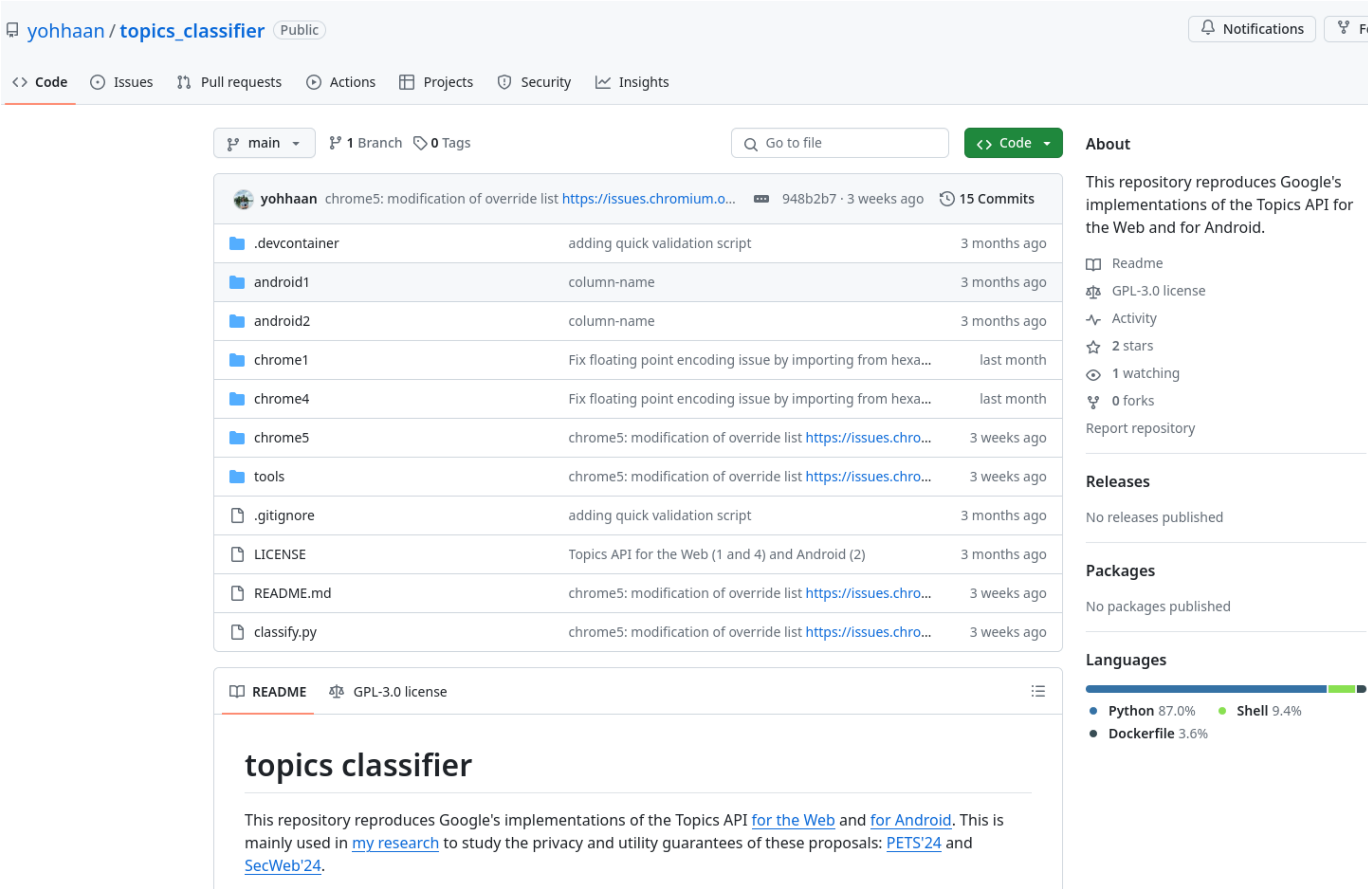
Ongoing Discussion & Contention

> **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.)

Topics Classifier



yohhaan / topics_classifier Public

yohhaan chrome5: modification of override list https://issues.chromium.o... 948b2b7 · 3 weeks ago 15 Commits

.devcontainer	adding quick validation script	3 months ago
android1	column-name	3 months ago
android2	column-name	3 months ago
chrome1	Fix floating point encoding issue by importing from hexa...	last month
chrome4	Fix floating point encoding issue by importing from hexa...	last month
chrome5	chrome5: modification of override list https://issues.chro...	3 weeks ago
tools	chrome5: modification of override list https://issues.chro...	3 weeks ago
.gitignore	adding quick validation script	3 months ago
LICENSE	Topics API for the Web (1 and 4) and Android (2)	3 months ago
README.md	chrome5: modification of override list https://issues.chro...	3 weeks ago
classify.py	chrome5: modification of override list https://issues.chro...	3 weeks ago

README GPL-3.0 license

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: PETS'24 and SecWeb'24.

About

This repository reproduces Google's implementations of the Topics API for the Web and for Android.

Readme
GPL-3.0 license
Activity
2 stars
1 watching
0 forks
Report repository

Releases

No releases published

Packages

No packages published

Languages

Python	87.0%
Shell	9.4%
Dockerfile	3.6%

https://github.com/yohhaan/topics_classifier

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

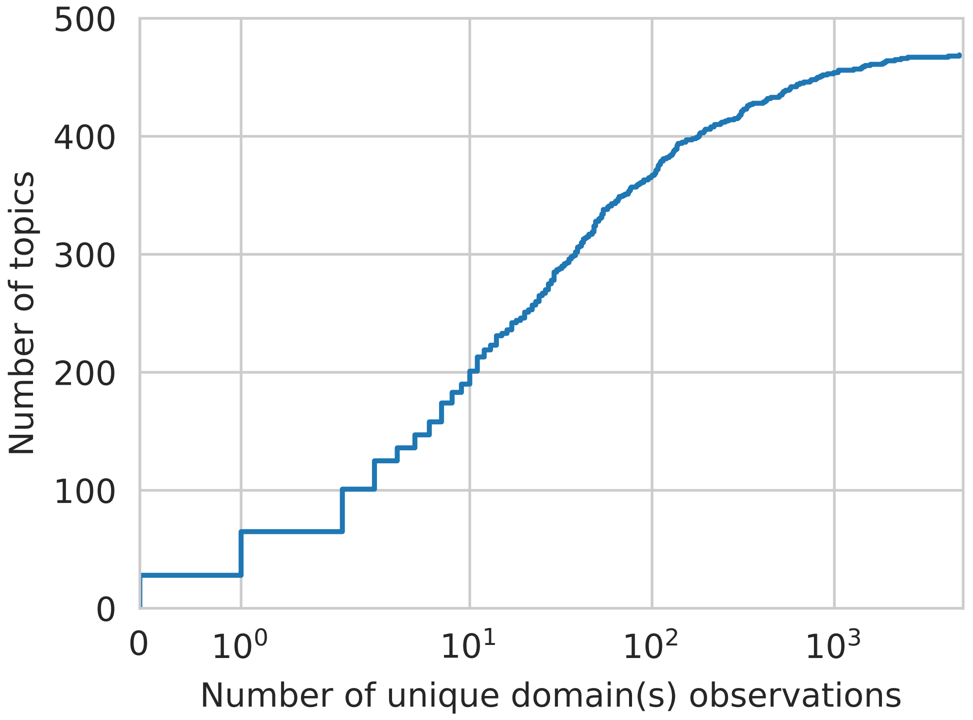
Weeks	1	2	3	4	5
Unique topics (469 topics)	219	216	220	221	226
Unique profiles (1 207 users)	1 127	1 132	1 142	1 143	1 154

Real Topics Profiles

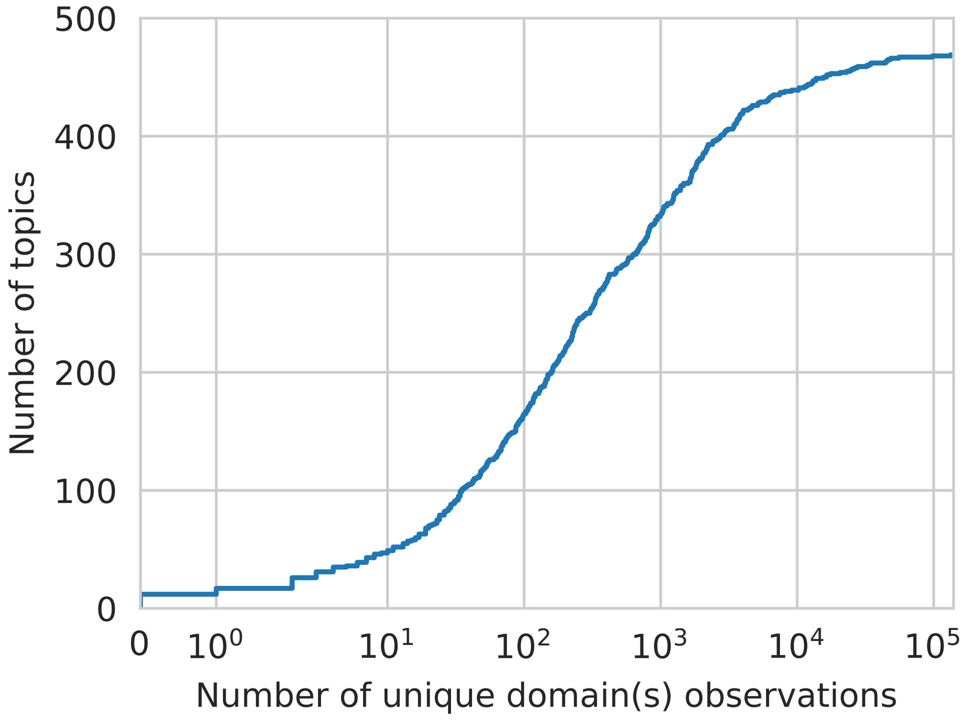
Stability

	Exactly 0	Exactly 1	Exactly 2	Exactly 3	Exactly 4	Exactly 5
From week 1 to 2	59 (4.9%)	187 (15.5%)	297 (24.6%)	369 (30.6%)	229 (19.0%)	66 (5.5%)
From week 2 to 3	70 (5.8%)	189 (15.7%)	318 (26.3%)	345 (28.6%)	226 (18.7%)	59 (4.9%)
From week 3 to 4	72 (6.0%)	188 (15.6%)	320 (26.5%)	325 (26.9%)	246 (20.4%)	56 (4.6%)
From week 4 to 5	70 (5.8%)	240 (19.9%)	324 (26.8%)	318 (26.3%)	211 (17.5%)	44 (3.6%)

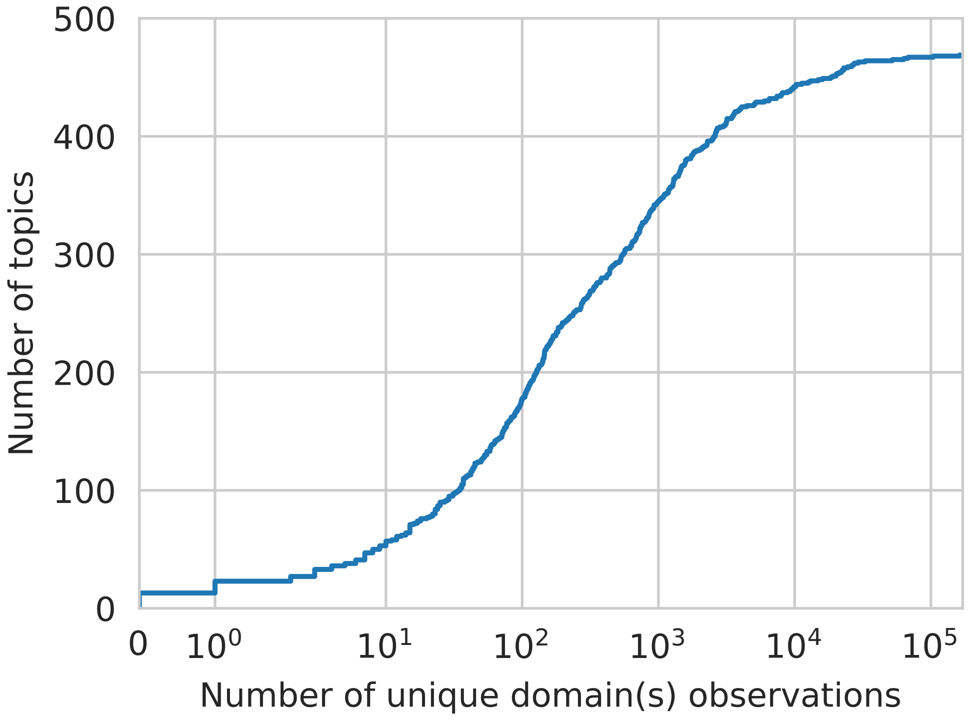
Noise Removal - Topics Distribution on the Web



Static Mapping



CrUX 1M



Tranco 1M

Noise Removal - Repetitions

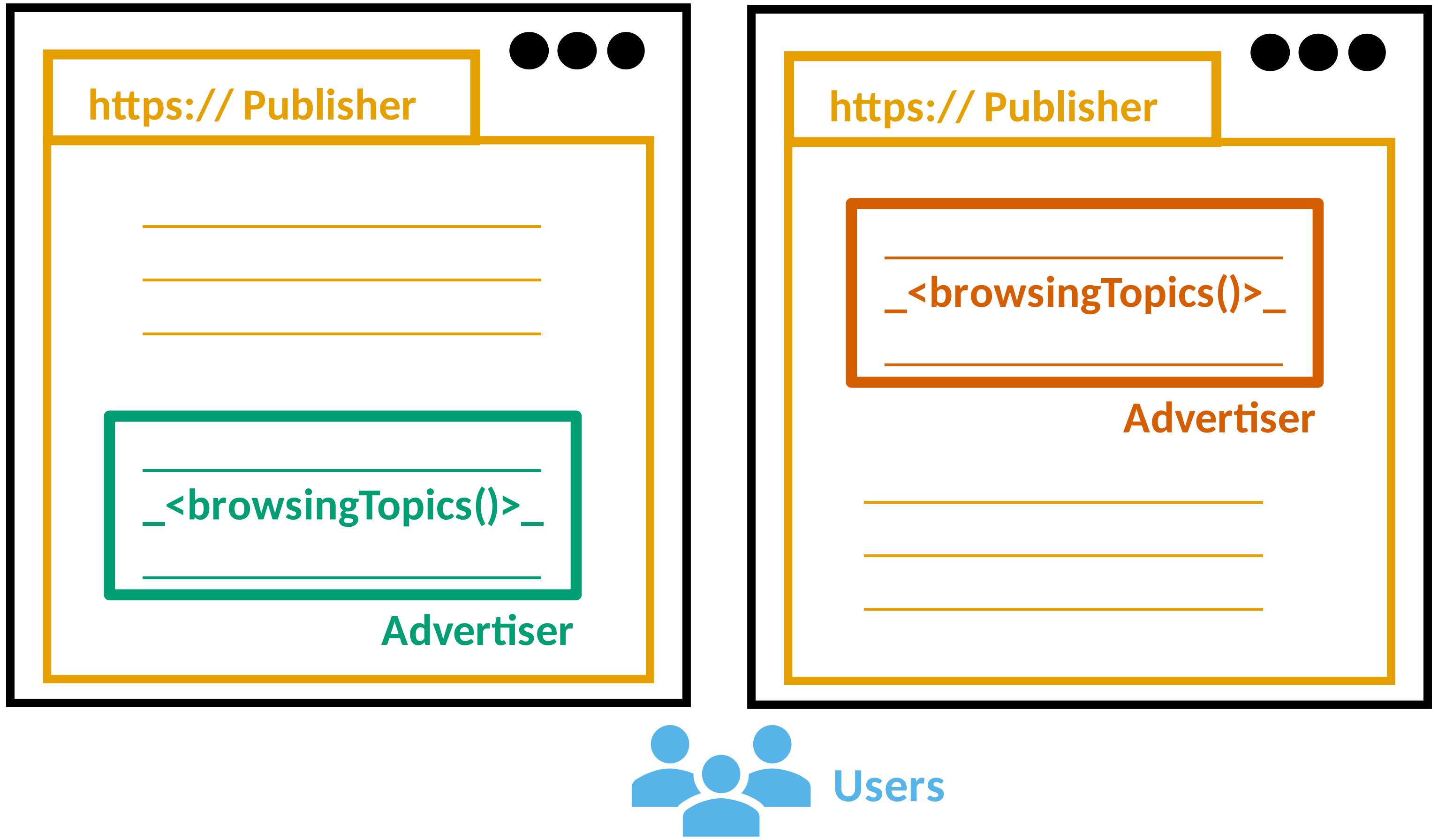
Epoch	Topics
0	✸, 📘, 📘
1	📘, 📘, ⛵
2	📘, ⛵, 🚴
3	⛵, 🚴, 🏏
4	🚴, 🏏, ⛵
5	🏏, ⛵, 🏏
6	⛵, 🏏, 🚴

✸ noisy
⛵, 📘, 🏏, 🎵, 🚴 genuine

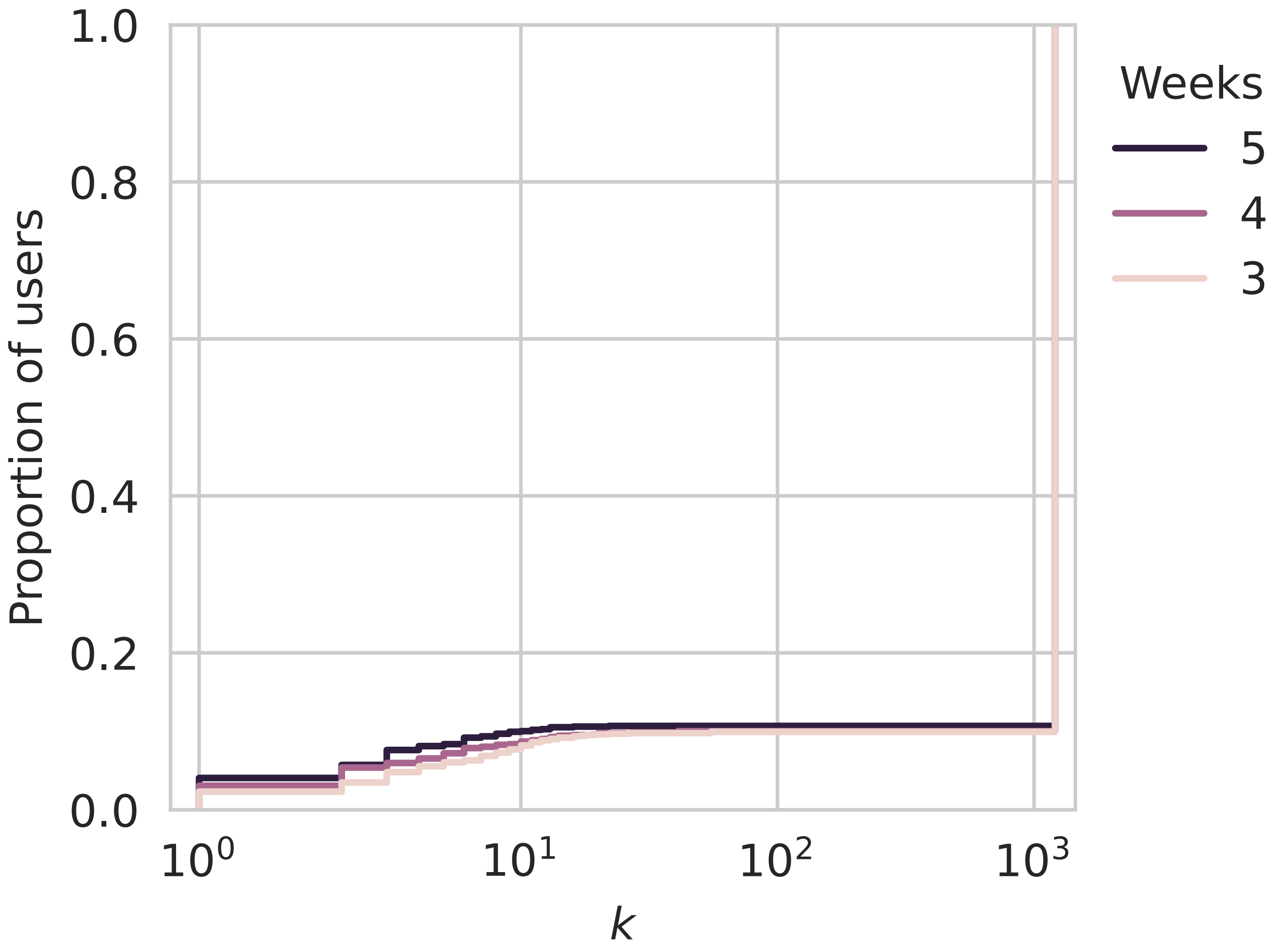
Simulation Results

Week	Accuracy	Precision	TPR	FPR
3	0.955	0.104	0.945	0.045
4	0.955	0.103	0.936	0.045
5	0.954	0.103	0.934	0.045

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

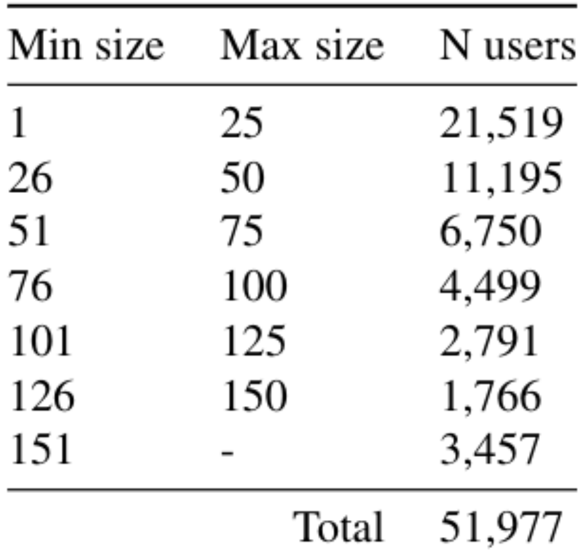
Additional Slides

Interest-disclosing Mechanisms for Advertising are Privacy-Exposing (not Preserving) PDF (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)



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

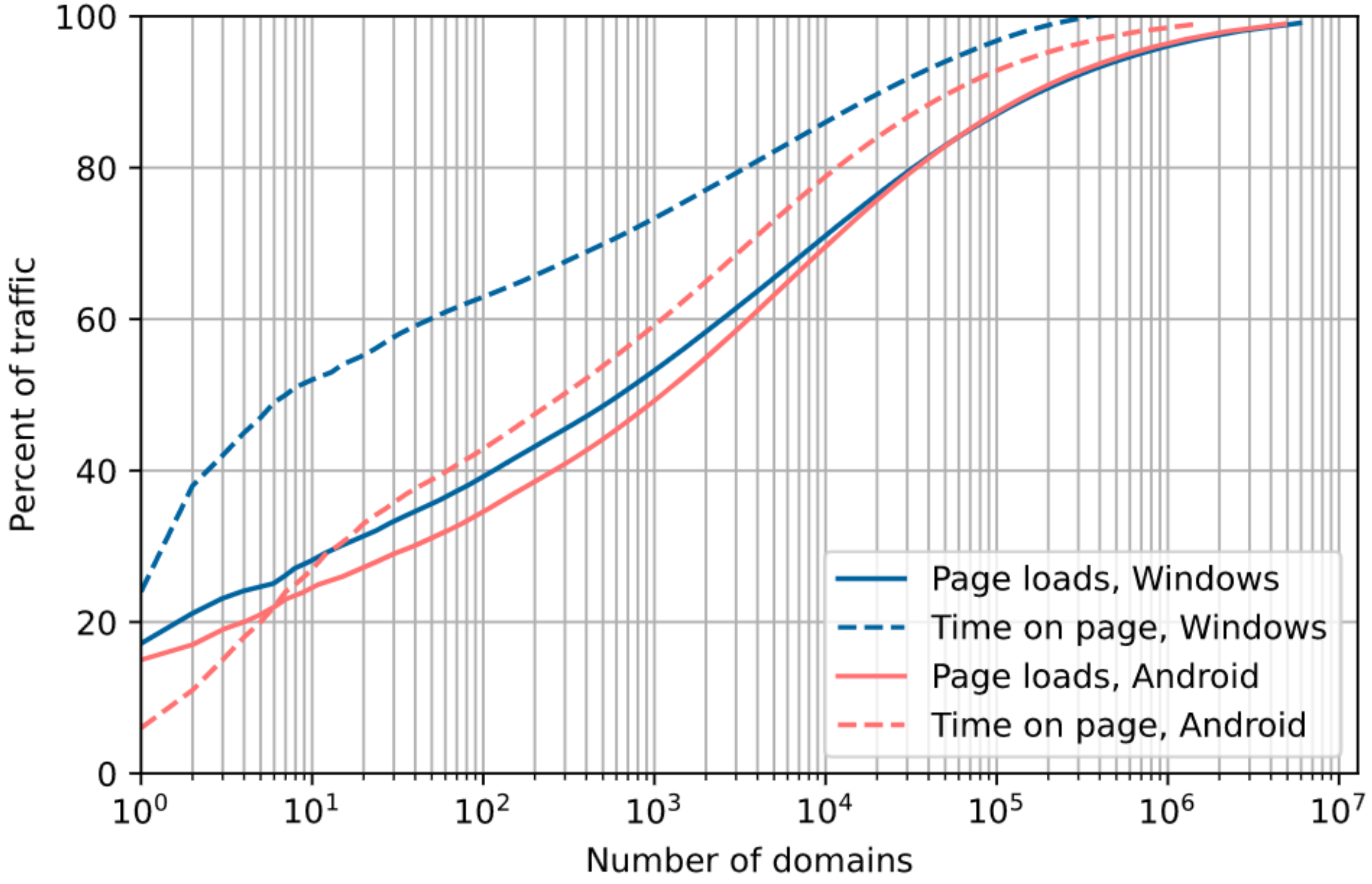
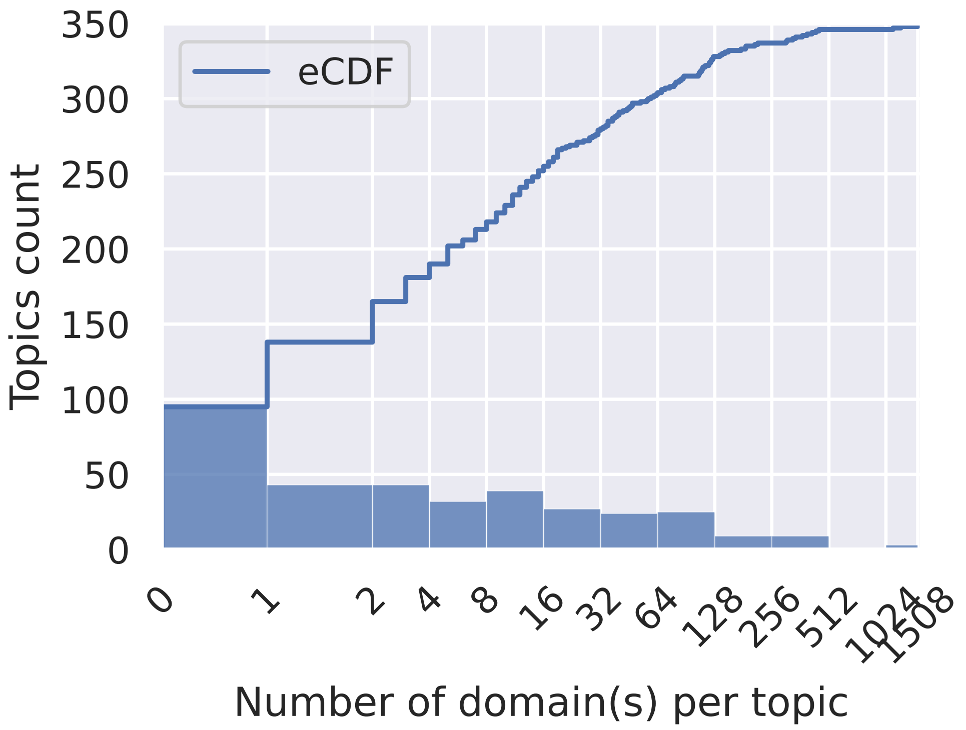
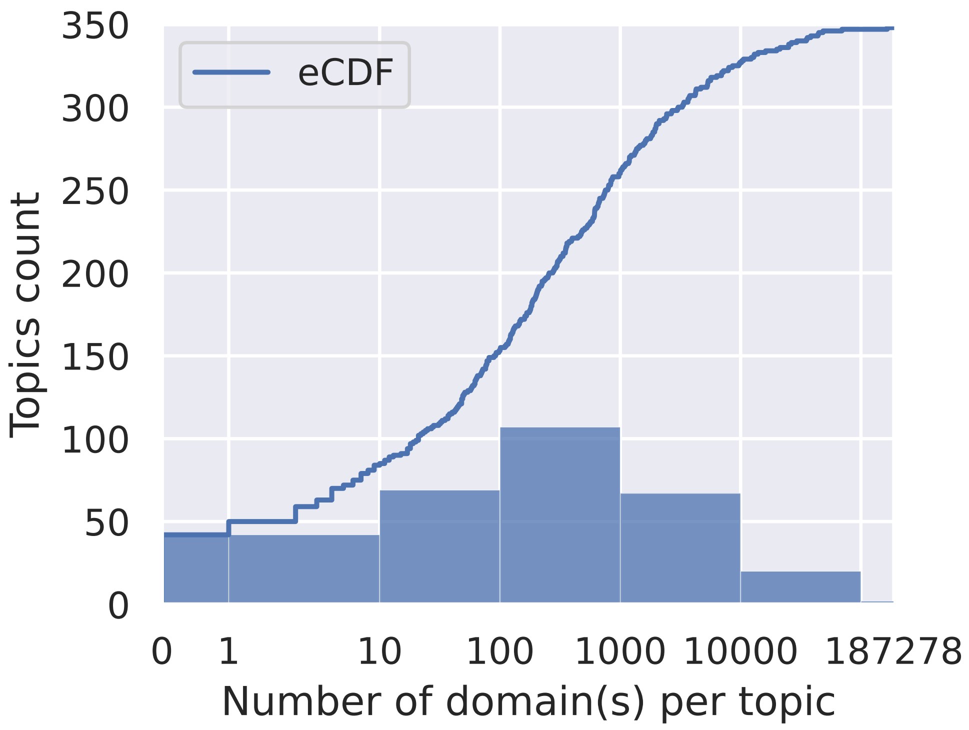


Figure 1: Distribution of Web Traffic By Website Rank

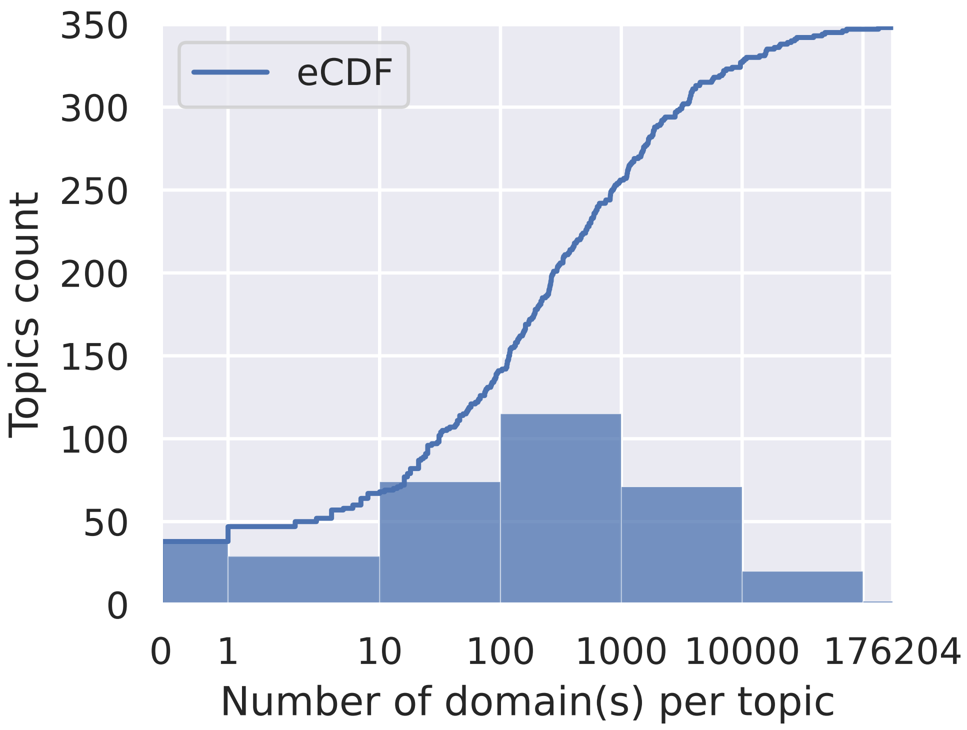
Noise Removal - Topics Distribution on the Web



Static Mapping



CrUX 1M



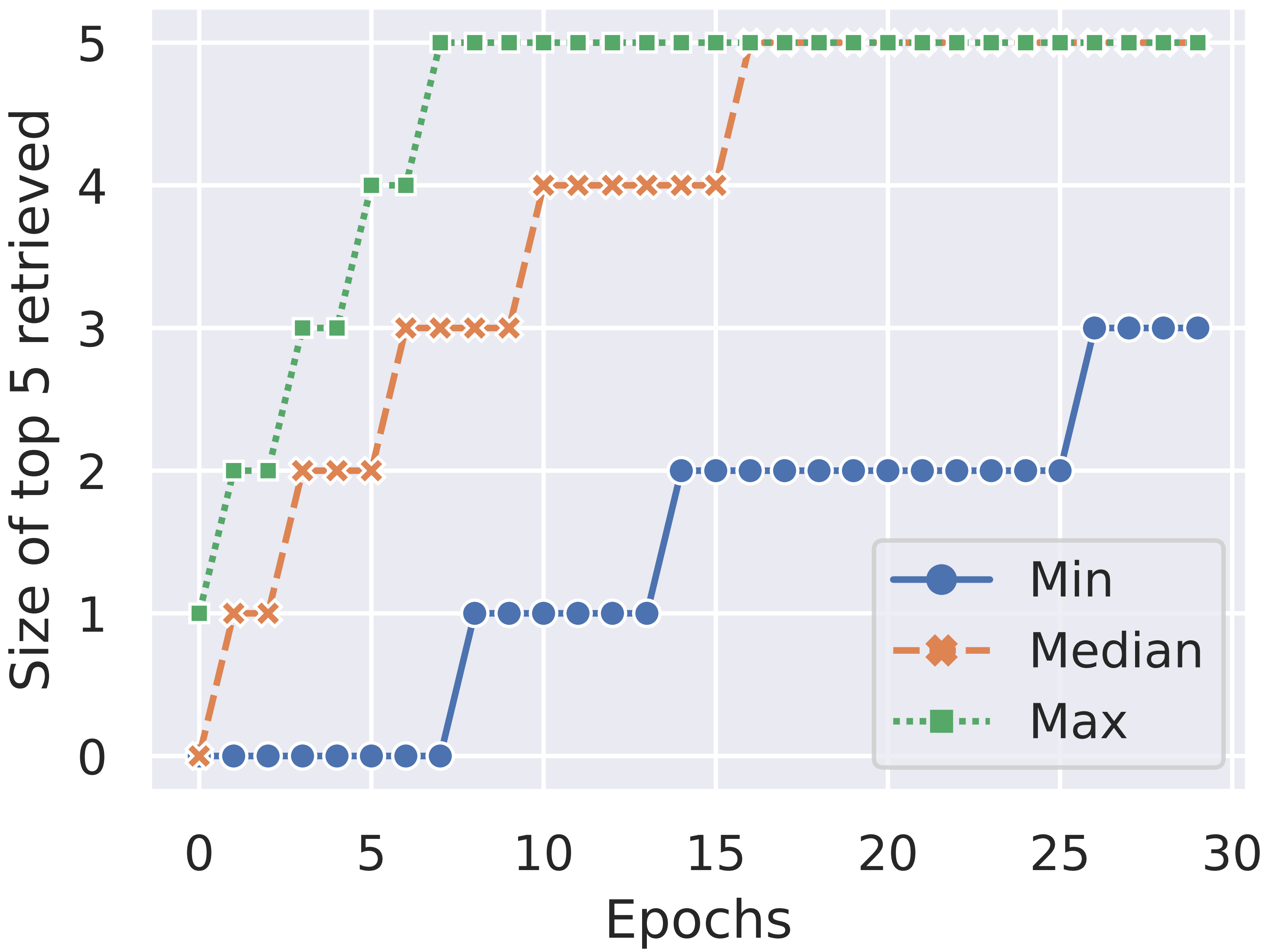
Tranco 1M

Noise Removal - Repetitions

Epoch	Topics
0	✱, 📖, 📖
1	📖, 📖, ⛵
2	📖, ⛵, 🚴
3	⛵, 🚴, 🏏
4	🚴, 🏏, ⛵
5	🏏, ⛵, 🏏
6	⛵, 🏏, 🚴

✱ noisy

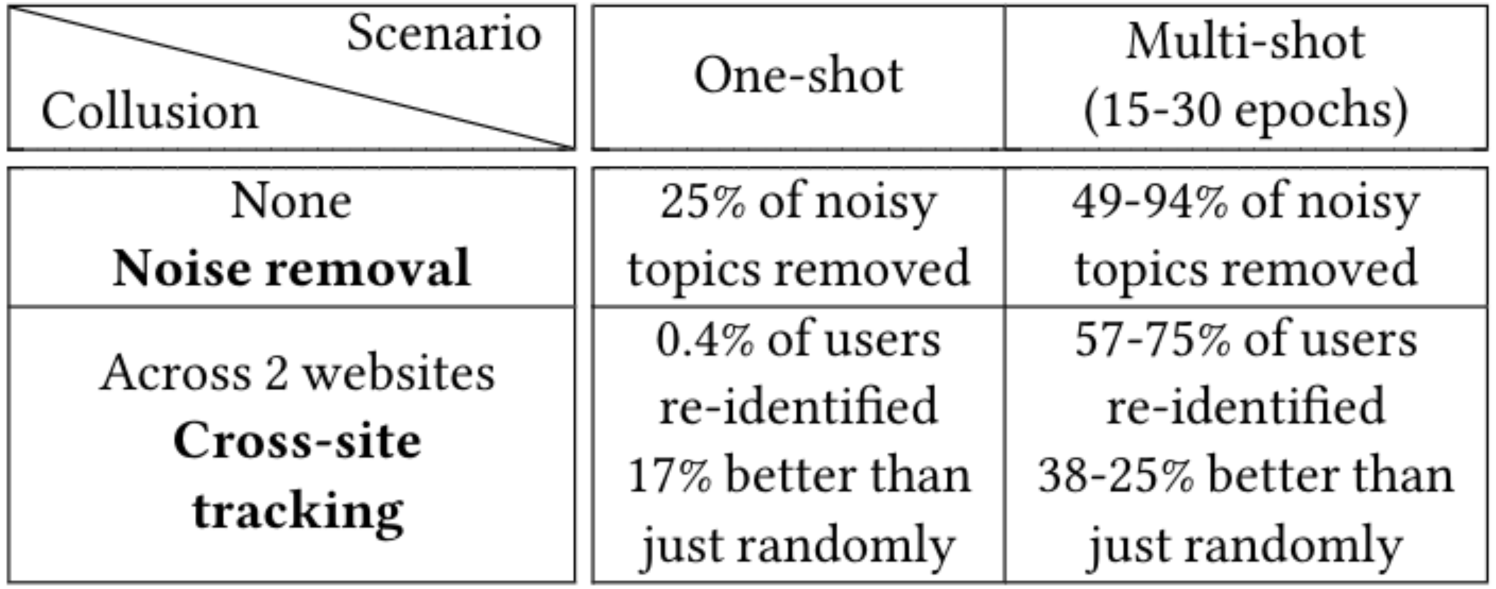
⛵, 📖, 🏏, ♫, 🚴 genuine



Top 5 Recovery

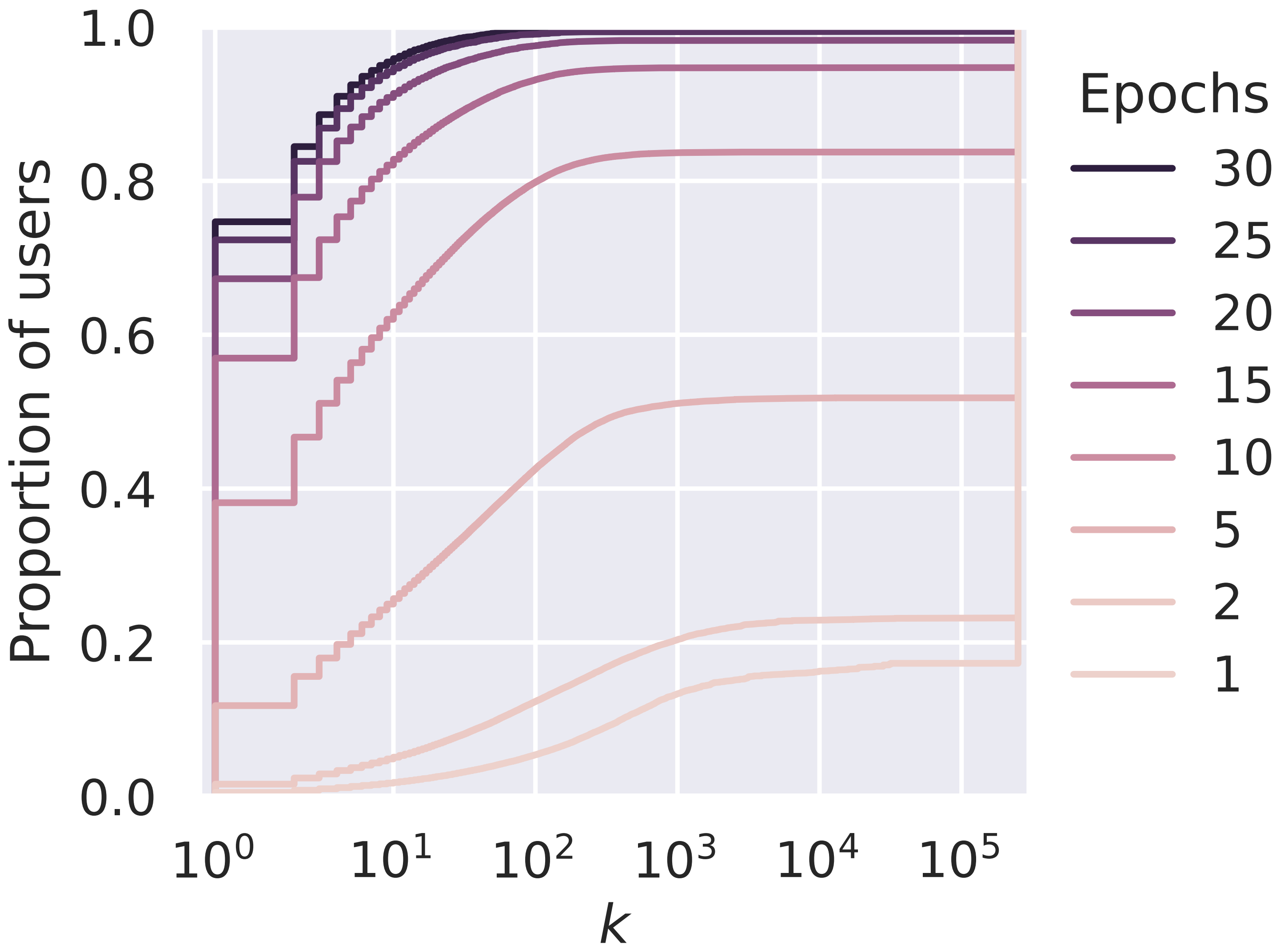
Advertisers can Re-identify Users

Simulation for 250k stable users



Scenario Collusion	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

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**

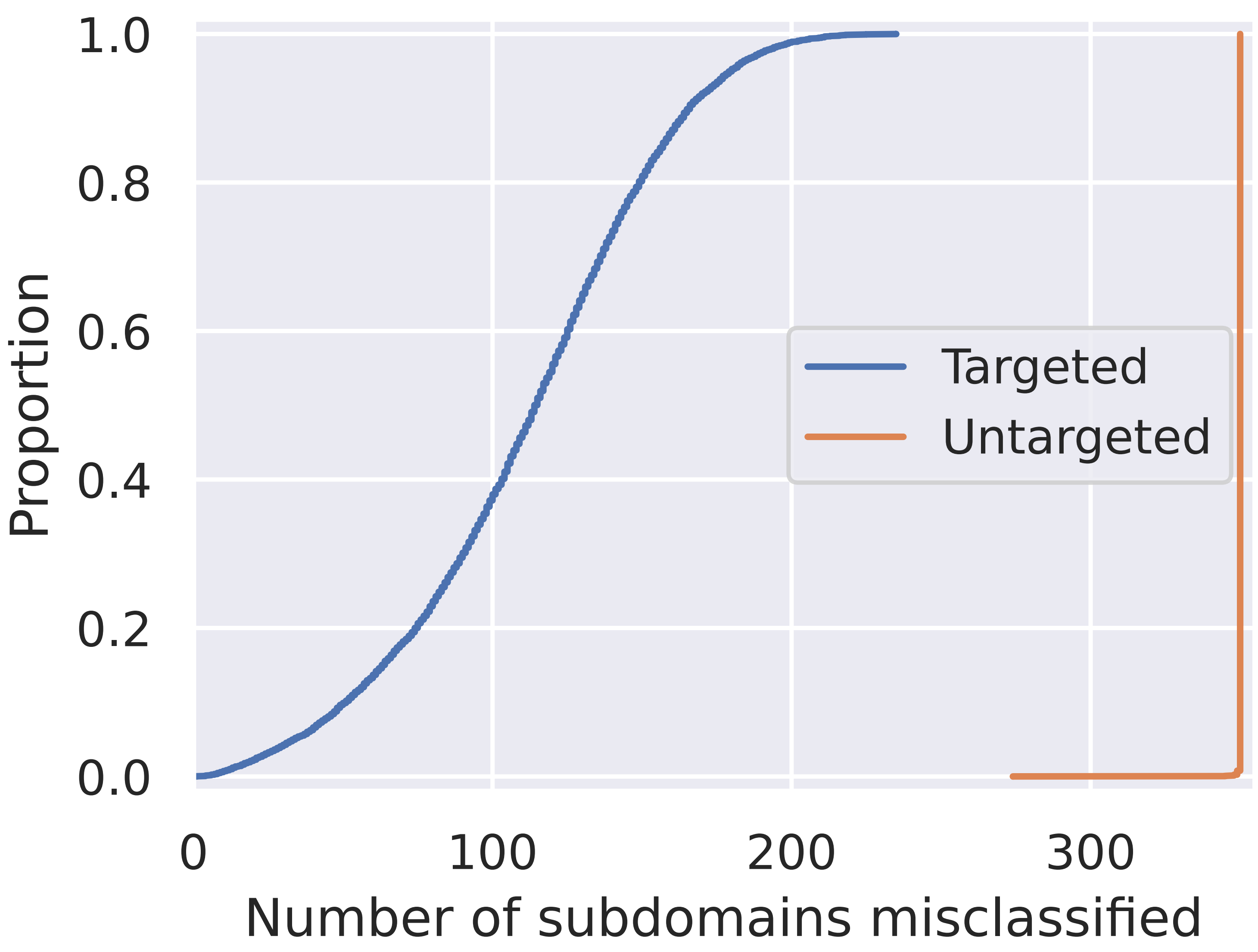
Misclassification

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


Proportion
Number of subdomains misclassified
Targeted
Untargeted