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.**

1. Introduction

The `Topics` API from `The Privacy Sandbox` is being developed by Google to deprecate third-party cookies on Chrome with the promise of both providing utility to advertisers and privacy to users [1]. `Topics` works by having the web browser classify the websites visited by users into categories of interest. Advertisers who are embedded on websites can observe some of the recent top users' topics and use that information to perform an ad auction [2], [3]. Prior analyses [4], [5], [6] have pointed at different limitations of the proposal. For instance, in a worst-case analysis quantifying the exact utility and privacy goals stated on the `Topics` proposal by Google, we demonstrated that users with stable interests have a higher risk of being re-identified across website visits. Indeed, over 60 % of the 250k stable users that we simulated are not guaranteed strictly more than 10-anonymity after 10 observations [6].

These results have led to an ongoing discussion between Google and research communities about the capability 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, and other modifications of the topics computation have been released since. Finally, we simulate how the `Topics` API would work for these users and evaluate if advertisers can fingerprint and re-identify users though their topics.

We find that each week at least 93 % of the users in this dataset have a unique top 5 topics profile. Measuring how stable these top profiles are over time, we obtain that at least 47 % of users have 3 or more topics in common in their top 5 topics profile from one week to the next one, while only less than 6 % of users have none. The combination of unique and stable profiles opens the possibility to use the `Topics` API to fingerprint users. We observe that in practice for the 1207 real users from the dataset, 46 %, 55 %, and 60 % of them are re-identified across 2 websites after 1, 2, and 3 observations of their topics by third parties, respectively. This paper highlights the importance of public and reproducible evaluations of any claim made by new web proposals and to identify the potential limitations of these techniques during their design rather than after their deployment.

1. Commit sha 78b13c3 as of February 19, 2024

We make the following contributions:

- The first privacy analysis of the newest version of the `Topics` API.
- The use of the largest publicly available dataset of real browsing histories to evaluate the `Topics` API.
- A reproducible methodology confirming on real data that users can be fingerprinted by their topics.

2. Background & Related Work

2.1. `Topics` API for the Web

Description. The `Topics` API for the web aims at deprecating third-party cookies while still enabling interest-based advertising [1], [2], [11]. Google proposes that without third-party cookies, web browsers be the ones to classify users' online behaviors into topics of interests and return the top visited topics to advertisers embedded on the websites being browsed. As of February 2024, the `Topics` API for the web is being rolled out to every Chrome users [12], [13], while other browsers such as Apple Safari [14], Mozilla Firefox [4], [15], or Brave [16] have declined to implement `Topics` in their product citing several privacy concerns they have.

Privacy Claims. Google claims that the `Topics` API for the web makes it *"difficult to reidentify significant numbers of users across sites using just the API"* and that *"the topics revealed [are] less personally sensitive about a user than [...] today's tracking methods"* [2].

Technical Details. Next, we summarize how the `Topics` API is implemented; we adopt the notations (see Appendix A) introduced in our prior work [6]. Figure 1 shows how at the end of an epoch e_0, the selection of the top $T = 5$ most visited topics for each user is performed; the websites visited by users are classified locally by their web browsers into topics from **the taxonomy** using the static mapping and the machine learning model released by Google. **The static mapping** is a list annotated by Google of domains to their corresponding topics, **the `Topics` classifier** checks first if the domain is present in that list before invoking the ML model for classification. Then, when an advertiser embedded on a website calls the `Topics` API, an array of maximum $\tau = 3$ topics and randomly shuffled is returned; for each of the last $\tau = 3$ epochs, either with probability $p = 0.05$, a random topic is drawn uniformly from the taxonomy, or with the opposite probability $q = 1 - p = 0.95$, a real one is picked from the corresponding top $T = 5$ most visited topics. **The witness requirement** in `Topics` enforces that a real topic is returned to API callers only if they were embedded on a website of the same topic that was also visited by the user during the past $\tau = 3$ epochs.

2.2. Analyses of the `Topics` API

Google's Analyses. Google released two privacy analyses of the fingerprinting risk of the `Topics` API; a white paper computing the aggregate information leakage per API call and for two consecutive calls [7] and a second work using a theoretical framework to measure re-identification risk [8]. However, the empirical measurements in both

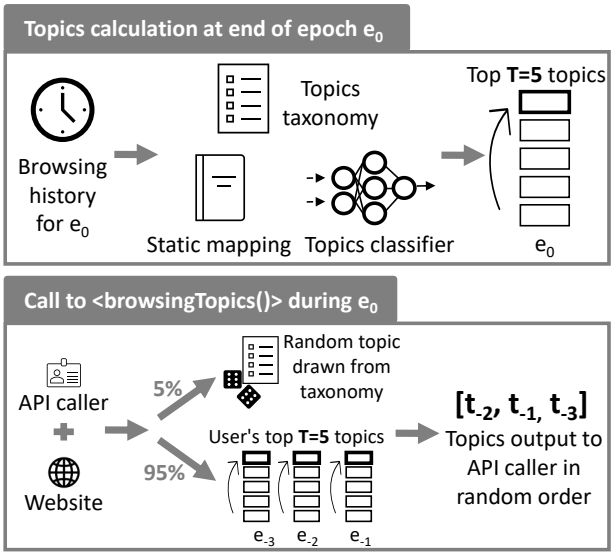
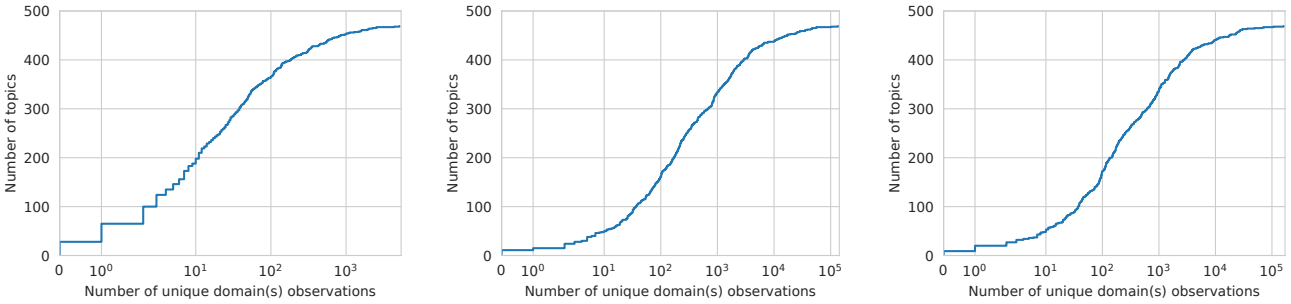


Figure 1: Overview of the `Topics` API for the web [6]

works have been performed on a private dataset, preventing the verification of the claims being made. Similarly, only aggregate and final results are reported, the lack of details across the distribution of users can hide the privacy risks of `Topics` for specific users as already pointed out by Thomson [4]. Additionally, Google's second analysis assumes and infers from an aggregate statistic that *"for every user, samples of top sets [of topics] are independent across time"* [8], while prior web measurement studies have found that users' interests exhibit some stability over time [17], [18], [19], [20], [21], [22], [23]. In this paper, we discuss these assumptions and methodologies, argue the need for reproducible evaluations of the `Topics` API, and perform one on real data.

Independent Analyses. One of the first analysis of the `Topics` proposal was carried out by Thomson, an engineer at Mozilla, who outlined potential limitations and risks associated with implementing the API [4]. Then, Jha et al., iterated on this initial analysis, collected browsing data on a sample of 268 users, and evaluated users' risk of being re-identified across websites through their topics of interest [5]. Concurrently, we performed a systematic evaluation of the `Topics` API for the web against Google's own stated goals, and uncovered how difficult it can be for the `Topics` API to deliver utility while guaranteeing privacy to all users [6]. In our worst-case analysis of the `Topics` API, we showed how advertisers could identify at least 25 % of the noisy topics added by the API to its results and how users with stable interests are at a higher risk to be fingerprinted across websites through their topics. We found that over 60 % of the 250k stable users we simulate are not guaranteed strictly more than 10-anonymity for 10 observations of their topics by advertisers. In addition to finding these limitations to Google's privacy objectives, we also measured the accuracy of the `Topics` classification, informing on the level of utility retained by advertisers with this new API, and demonstrated how third parties can abuse the system by causing misclassifications [6].



(a) Static mapping (b) CrUX (c) Tranco

Figure 2: Empirical cumulative distributions of the number of topics per number of unique domain(s) as observed on different lists of most visited domains.

3. Methodology

In the following section, we present the modifications that have been made to the `Topics` API since previous analyses were published and describe the dataset of real browsing histories on which we simulate `Topics`.

3.1. New `Topics` API Version

Since the last analyses of the `Topics` API [4], [5], [6], Google made a few modifications to the implementation to improve its utility [24], we list them here, the versions used in this paper are also listed in Appendix A.

Modifications to Domains Classification. First, a new taxonomy (version *2*) of $\Omega = 469$ topics has been released: 160 topics from the initial taxonomy have been removed and 280 new ones added. The model classifier (version *4*) has been updated accordingly, and the static mapping that consists of a list of websites to topics manually annotated by Google has grown from about 10k to now 50k domains.

Modifications to Top Topics Selection. In the newest version of the `Topics` API, the selection of the top $T = 5$ topics for each epoch has been modified to not only take into account the frequency of each topic, but also their utility for ad relevance. Each topic from the taxonomy is now assigned a *standard* or *higher* utility label, and top topics are now selected by first ranking topics by utility and then frequency. This modification was made by Google after receiving feedback from advertisers that topics observed many times and shared by a lot of users (general topics like *News* or *Arts & Entertainment*) were not so useful for advertising purposes [24]. As a result, in the new version, topics of higher utility (less general, but also potentially more revealing privacy-wise) are considered before others.

Re-implementation and Topics Distribution. For this paper, we re-implement the latest version of the `Topics` API (commit sha 78b13c3) shipped in Google Chrome locally to be able to classify several thousands domains at once. This lets us obtain the prior distributions of topics observations on the most visited websites; Figure 2 shows such distributions on the following three lists of domains: the new static mapping used by Google composed of about 50k domains (Figure 2a), and the top 1M most visited domains as reported by CrUx [25] (Figure 2b) and Tranco [26] (Figure 2c), respectively. While the `Topics` API has been updated since prior analyses, results about the asymmetry of the observations of topics from the new taxonomy on the most visited websites still hold [6]; 28, 11, and 9 topics never appear at all from the taxonomy for the static mapping, CrUX, and Tranco, respectively, others only a few times, and very few on many domains from each list. Such prior distribution can be used by third party to distinguish if an observed topic is a noisy one (picked randomly from the taxonomy by the API in $p = 5\,\%$ of the cases) if it never appears or only seldomly on the top 1M domains in practice (see Section 4.4 for more).

3.2. Dataset of Real Browsing Histories

Initial Dataset. In this paper, we use data collected in October 2018 on 2148 users from Germany who have volunteered to share their desktop browsing histories in exchange for a financial compensation. The dataset has been shared online by researchers studying the habits [9], [10] and uniqueness [27] of web users activity. It was anonymized by only releasing the fully qualified domain name of the visited websites and not the complete URL of the pages browsed; thus avoiding to make public social media identifiers, personal profile pages, etc. [28]. Overall, these 2148 users visited 9 151 243 URLs across 49 918 unique domains or 67 300 unique origins during 5 weeks.

Top Topics Profiles. We classify all the origins visited by the users during the 5 weeks from the initial dataset with the `Topics` classifier. This allows us to generate the top interests profiles of each user, exactly like the `Topics` API would do in practice if it were run on these browsing histories. We keep only users that have had a browsing history (and so topics assigned to them) for every week. Finally, we obtain a dataset composed of 1207 users who have visited 7 746 193 URLs across 43 684 unique domains or 58 370 unique origins in October 2018.

4. Privacy Evaluation

In this section, we perform a privacy analysis of the latest version of the `Topics` API on a dataset of real browsing histories. We start by outlining the privacy and utility trade-off of the `Topics` API.

4.1. Privacy-Utility Trade-off

Utility. The `Topics` API discloses to advertisers the interests that users had in the past epochs. Using these interests, third parties can infer which ads to display to users. Thus, the utility of the `Topics` API for advertisers is determined by how aligned the returned topics are with users' true interests. This is impacted by several factors such as the granularity of the taxonomy of topics, the accuracy of the classification of users' web behaviors by the API, the number of topics returned per API call and their nature ($5\,\%$ of the topics observed are expected to be noisy topics), the length of an epoch, as well as the number of websites on which the advertiser is embedded.

Privacy (k-anonymity). Because the `Topics` API prevents advertisers from assigning unique identifiers to users through third-party cookies, Google claims that `Topics` is better for online users [2]. `Topics` mainly relies on *k-anonymity*; the idea is that advertisers would observe several users with the same topics of interest during an epoch, and so would not be able to distinguish and track them. In practice, `Topics` still discloses information about users to advertisers and one concern is that third parties use these signals for tracking. Users with stable and unique interests have a higher risk of being re-identified through their topics [4], [6]. Indeed, although the API only discloses a maximum of 1 new topic per consecutive epoch after the initial call, advertisers can recover over time users' stable interests. These interest profiles collected over a longer period of time contain more information than just a single API call, and if they are unique enough they can be used to fingerprint users. Thus, in Section 4.3, we measure how stable and unique real users' interests are before evaluating their deanonymization risk in Section 4.5.

Privacy (Plausible Deniability). Google also claims that `Topics` does not disclose sensitive information about users because topics related to sensitive categories (e.g., religion, politics, adult themes, etc.) have been removed from the taxonomy [2]. This assumes that no topic in the current taxonomy can be associated with any sensitive subject which is not guaranteed; correlations between interests or within a specific context can be made [4], [14], [15]. Additionally, `Topics` aims to provide plausible deniability to users by adding random topics to $5\,\%$ of the results, so that users could blame on this randomness and ambiguity a topic observed by an advertiser rather than disclosing it was part of their browsing behavior [2]. To improve the accuracy of their ads, advertisers will logically want to distinguish between noisy and true topics. We already established in Section 3.1 that the distribution of topics on the web was still not uniform at all. In Section 4.4, we re-evaluate on real data if this can be leveraged to identify noisy topics as in prior work [6].

Trade-off. As a result, there is fundamental trade-off between privacy and utility within the design of the `Topics` API; by disclosing users' interests directly to advertisers, the API also discloses some information about their browsing behaviors that can in some cases be used for tracking. Perhaps, the best illustration of this trade-off is the recent introduction of utility label along each topic of the taxonomy and the modification to how top topics are computed. Advertisers reported that during the origin trials of the `Topics` API, the results they were observing were too generic in scope and shared by many users (like *News* or *Arts & Entertainment*). Because they were not so valuable to advertisers, Google modified the `Topics` API to output first topics with high utility rather than high frequency [24]. If these changes aim to address utility issues faced by advertisers, they also impact users' privacy as the new topics now returned are potentially more unique and shared by less users.

4.2. Analysis

Research Questions. We ask the following questions:

(Q1) How stable and unique are real users' profiles?
(Q2) Can we flag noisy topics returned by real users?
(Q3) Can we track across websites these real users?

Assumptions. To evaluate the privacy guarantees provided by the `Topics` API to these real users, we assume:

- No witness requirement applied to API callers, i.e., we consider that they can observe any topic. Note that this is aligned with the previous analyses of `Topics` including Google's [4], [5], [6], [7], [8].
- Advertisers are observing and recording the topics of the same users across consecutive weeks. This lets them infer which topic likely corresponds to which week and so, we do not randomize the array returned by the API in our simulation.

Open Source Artifact. To enable replication of our results and further analyses, we release our code available at https://github.com/yohhaan/topics_api_analysis.

4.3. Stability and Uniqueness of Topics Profiles

We seek to answer **(Q1)** by measuring how unique and stable the top topics profiles of the real users in our dataset are. Indeed, prior work showed that users with stable and unique interests in a population are at more risk of being re-identified by participating in the `Topics` API [4], [6]. Table 1 reports how stable the top $T = 5$ topics profiles of the real users in our dataset are across weeks. We can see from these results that very few users have top topics profiles totally unstable from week to week, but that rather users exhibit similar interests over time; less than $6\,\%$ of the users have 0 topic in common, while at least $47\,\%$ have 3 or more stable interests over time. Table 2 shows how many unique topics from the taxonomy are observed among all users each week; on this dataset of real browsing histories from 2018, only about 220 out of the 469 topics of the current taxonomy are present in the true topics of interest of users. Table 2 also reports for each week how many users have unique top $T = 5$ topics profile; we observe that almost all of the 1207 users have unique topics profiles in this dataset. These results are aligned with prior ones in web measurement studies finding that web behaviors and interests of real users are unique [29], [30], [31], [32] and follow a recurrent system [17], [18], [19], [20], [21], [22], [23]; users repeat actions leading to similar interests across weeks while also performing new activities. Thus, worst-case analyses assuming users' interests to be stable are important to

TABLE 1: Number (and proportion) of users having 0 to 5 topics stable in their top 5 profiles for consecutive weeks.

	0 topic in common	Exactly 1 topic	Exactly 2 topics	Exactly 3 topics	Exactly 4 topics	Exactly 5 topics
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%)

TABLE 2: Unique topics and top profiles across weeks.

Week	1	2	3	4	5
Unique topics	218	215	220	223	226
Unique profiles	1132	1132	1144	1145	1151

understand the privacy limitations of the `Topics` API. Next, we adapt the privacy analysis of our prior work [6] to measure these limitations on real data.

4.4. Identifying Noisy Topics

To address **(Q2)**, we hypothesize that some of the noise added $p = 5\,\%$ of the times by the API to the real topics observed by advertisers–in order to provide plausible deniability to users–can be removed by following a similar approach than in our prior work [6]. This consists for advertisers in flagging topics that are never or seldomly observed on the web, but returned to them by the `Topics` API as noisy. Similarly, if advertisers determine that a topic is observed repeatedly for a user across different corresponding epochs, the topic is very likely to be a true one by the probabilistic nature of how the topics returned are picked. To decide if a topic can be naturally observed on the web or not, we use the distribution of topics on the top 1M most visited websites given by CrUX that we classified in Section 3.1. We set a threshold of 10 domains on which each topic needs to appear to not be classified as noisy. This threshold corresponds to the one picked in our previous work [6], but we intend in future work to explore different thresholds and other approaches to try to identify noisy topics on this dataset of real browsing histories. We then simulate for each user and for each week 100 calls to the `Topics` API, flag topics we believe to be noisy, and compare the classification to the ground truth. We report in Table 3 the accuracy, precision, true positive rate (TPR) and false positive rate (FPR) of the classification of these 100 observations per user and per week (the positive and negative classes of our classifier are *noisy* and *real* topic, respectively). Note that the `Topics` API only outputs a noisy topic in about $5\,\%$ of the cases, as such this experiment presents a class imbalance, which is why we will mainly look at the *precision* metric; the proportion of correctly flagged noisy topics among all instances flagged as noisy. Additionally, a topic observation requires knowing a user's top interests for the last $\tau = 3$ weeks which explains why the table starts on week 3. We find that with this approach, $10\,\%$ (precision) of the topics we flag are truly noisy which is better than $0\,\%$ if we were considering all topics to be genuine users' interests or the expected $5\,\%$ (remember the class imbalance) of a random guess. Thus, it is still possible in practice and with the latest version of `Topics` to refute plausible deniability to users for some topics. We leave to future work other classification heuristics.

TABLE 3: Classification results of the identification of the noisy topics returned by `Topics`.

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

4.5. Re-identifying Users Across Sites

Finally, to respond to **(Q3)**, we run the following re-identification experiment [6], [8]. We consider that each user visits two different websites on which two different advertisers that are colluding are embedded. For each week that we simulate, these advertisers are observing the topics of each user from their respective vantage points. Their objective is to correlate a user seen by one advertiser on the first website with the correct user seen by the other advertiser on the second website. For each user, this matching is done by returning the user(s) that have the lowest Hamming distance between the topics observed by the second advertiser and the set of topics observed by the first advertiser [8]. We leave to future work the exploration of other matching strategies. Also, this re-identification is performed every week with only the topics that have been observed until that week for each user. We present the results of this experiment in terms of level of k-anonymity that is provided to users by the API; i.e.,, $k = 1$ means that advertisers have been able to uniquely re-identify the user across their visits for that specific week, $1 < k < n$ where n is the number of users means that the user's visits could not be uniquely correlated to each other, but with the visits of $k - 1$ other users, and $k = n$ means that the re-identification failed and is no better than a random uniform guess. Figure 3 shows the cumulative distribution function of the proportion of users that are satisfied k-anonymity (log scale) across the simulated weeks (recall that a topic observation for a user requires knowing their top interests for the last $\tau = 3$ weeks, thus we start at week 3). Note that the experiment is also valid, and our results exactly the same, if we were to consider instead of two advertisers only one embedded on both websites and trying to correlate visits from both vantage points. We observe that $46\,\%$, $55\,\%$, and $60\,\%$ of the users are uniquely re-identified across their visits after only 1, 2, and 3 observations, respectively. For the 3 weeks simulated, we observe that about $65\,\%$ of the users are mapped to a group of size $k < n = 1207$, which means that their cross-site visits can be correlated to each other with a higher probability than just a random guess. We conclude that an important part of the users from this real dataset are re-

identified across websites through only the observations of their topics of interest in our experiment. Thus, the real users from our dataset can be fingerprinted through the `Topics` API. Moreover, as can be seen, the information leakage and so, privacy violation worsen over time as more users are uniquely re-identified.

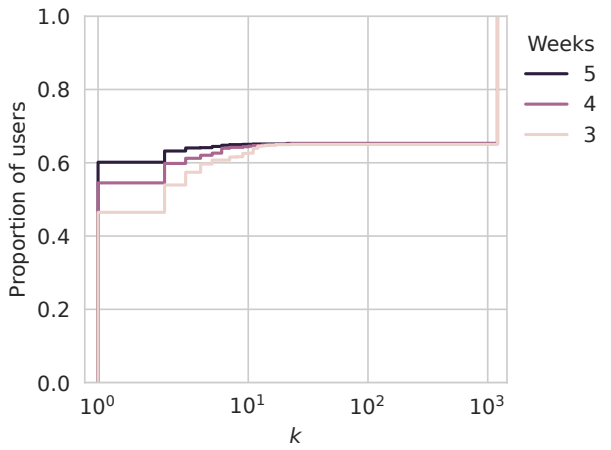


Figure 3: Cumulative distribution function of the proportion of users from the dataset that are provided k-anonymity. Note that $k = 1$ means that users are uniquely re-identified and so fingerprinted by the API.

5. Discussion

Limitations. The dataset of real browsing histories used in this paper likely contains some biases; it was collected in Germany in October 2018 and so does not necessarily represent every individual browsing the web today. Additionally, its size remains somewhat limited in number of users or collected weeks. Nevertheless, the use of this dataset does inform us on the privacy guarantees that `Topics` would have delivered in practice to these specific real users if the API had been available at the time the data was collected. This is perhaps the most relevant in our endeavor to address the need to trust Google's assertions and results made on their private dataset. Indeed, this is, to the best of our knowledge, the largest publicly available dataset of real web histories on which a reproducible privacy evaluation of the `Topics` API has been performed.

Recommendations for Reproducible Evaluations. First, in order to evaluate the privacy and utility guarantees of a proposal, its goals should be clearly defined to be provable or quantifiable; in prior analysis, we had to rephrase the initial objectives stated by Google on the `Topics` API proposal as they were too vague [6]. Second, one should avoid performing non-reproducible evaluations on a private dataset [7], [8] as it limits the scope of these analyses to having to trust assertions being made without the possibility to verify them. If an evaluation on a private dataset needs to be performed, we would recommend publishing rigorous details about the experiment and the results rather than just final and aggregated metrics; indeed, Thomson already demonstrated the danger of reporting the average case rather than the full distribution on the privacy claims of the `Topics` API [4], [15]. To solve this problem, we would like to call on web actors like Google–that have access to real browsing histories–to release the topics of interest of a representative population of users across several months, so that it can serve as a common starting point for future analyses. Note that such dataset not necessarily needs to include the browsing histories of users, but just the set of top topics of anonymized users. It could even be synthetically generated by drawing these topics profiles from a representative distribution of real users to address potential privacy concerns by releasing real data.

Future Work. Overall, rigorous and thorough analyses of these new web proposals (`The Privacy Sandbox` for the web now encompasses more than 30 different mechanisms [33]) are needed; several interesting questions remain unsolved, such as if these proposals can deliver on their guarantees to all users, if their specification and implementation are correct, if they can be manipulated or abused, if their results can be correlated with other fingerprinting signals, if developers and users do understand how these new APIs work, etc. Regarding the `Topics` API, evaluating it without having access to Google's private dataset of browsing histories can be quite challenging; we refer readers to previous analyses of the `Topics` API to see how prior authors (partially) got around that obstacle [5], [6]. In this paper, we call on web actors to release anonymized or synthetic datasets to help with that point. In the meantime, we envision to perform a literature review and a taxonomy of web users measurement studies, pair these results with the few datasets of browsing histories publicly available, and generate synthetic distributions of users' topics of interest.

6. Conclusion

This paper complements previous analyses of the `Topics` API for the web by confirming on real data that the proposal does not provide the same privacy guarantees to all users. Specifically, we have shown on this real dataset of 1207 users that the `Topics` API can be used to fingerprint 60 % of these users across websites after 3 observations. We also highlighted the need for reproducible analyses of the claims made by new web proposals before their deployment. This is important to avoid replicating the same mistakes already made in the past with the same technologies being deprecated today; the introduction of e.g., web cookies later extensively abused for tracking purposes. As such, we call on web actors and industry to release representative and anonymized datasets to enable reproducible evaluations of the `Topics` API and other proposals from `The Privacy Sandbox`.

Acknowledgments

We sincerely thank the reviewers for their constructive comments and suggestions on this paper.

Funding acknowledgment: This material is based upon work supported by the National Science Foundation under Grant No. CNS-232088 and Grant No. CNS-2343611. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] Google, "Topics API: Relevant Ads without Cookies - The Privacy Sandbox," 2022. [Online]. Available: https://privacysandbox.com/proposals/topics/

[2] ——, "GitHub - patcg-individual-drafts/topics: The Topics API," 2022. [Online]. Available: https://github.com/patcg-individual-drafts/topics/

[3] ——, "Topics API developer guide," Apr. 2023. [Online]. Available: https://developer.android.com/design-for-safety/privacy-sandbox/guides/topics

[4] M. Thomson, "A Privacy Analysis of Google's Topics Proposal," Mozilla, Tech. Rep., Jan. 2023. [Online]. Available: https://mozilla.github.io/ppa-docs/topics.pdf

[5] N. Jha, M. Trevisan, E. Leonardi, and M. Mellia, "On the Robustness of Topics API to a Re-Identification Attack," *Proceedings on Privacy Enhancing Technologies*, 2023. [Online]. Available: https://petsymposium.org/popets/2023/popets-2023-0098.php

[6] Y. Beugin and P. McDaniel, "Interest-disclosing Mechanisms for Advertising are Privacy-Exposing (not Preserving)," *Proceedings on Privacy Enhancing Technologies*, 2024. [Online]. Available: https://petsymposium.org/popets/2024/popets-2024-0004.php

[7] A. Epasto, A. M. Medina, C. Ilvento, and J. Karlin, "Measures of cross-site re-identification risk: An analysis of the Topics API Proposal," p. 12, 2022.

[8] C. J. Carey, T. Dick, A. Epasto, A. Javanmard, J. Karlin, S. Kumar, A. M. Medina, V. Mirrokni, G. H. Nunes, S. Vassilvitskii, and P. Zhong, "Measuring Re-identification Risk," Apr. 2023. [Online]. Available: http://arxiv.org/abs/2304.07210

[9] J. Kulshrestha, M. Oliveira, O. Karacalik, D. Bonnay, and C. Wagner, "A web tracking data set of online browsing behavior of 2,148 users," Dec. 2020. [Online]. Available: https://zenodo.org/records/4757574

[10] ——, "Web Routineness and Limits of Predictability: Investigating Demographic and Behavioral Differences Using Web Tracking Data," Dec. 2020. [Online]. Available: http://arxiv.org/abs/2012.15112

[11] S. Dutton, "Topics API: developer guide," Jan. 2022. [Online]. Available: https://developer.chrome.com/docs/privacy-sandbox/topics/

[12] Google, "Preparing to ship the Privacy Sandbox relevance and measurement APIs," May 2023. [Online]. Available: https://developer.chrome.com/blog/shipping-privacy-sandbox/

[13] ——, "Shipping the Privacy Sandbox relevance and measurement APIs," 2023. [Online]. Available: https://developers.google.com/privacy-sandbox/blog/privacy-sandbox-launch

[14] A. van Kesteren, "The Topics API · Issue #111 · webKit/standards-positions," 2022. [Online]. Available: https://github.com/webKit/standards-positions/issues/111#issuecomment-1359609317

[15] M. Thomson, "Request for Position: Topics API · Issue #622 · mozilla/standards-positions," 2022. [Online]. Available: https://github.com/mozilla/standards-positions/issues/622

[16] P. Snyder, "Google's Topics API: Rebranding FLoC Without Addressing Key Privacy Issues," Jan. 2022. [Online]. Available: https://brave.com/web-standards-at-brave/7-googles-topics-api/

[17] S. Greenberg, *The Computer User as Toolsmith: The Use, Reuse and Organization of Computer-Based Tools*, ser. Cambridge Series on Human-Computer Interaction. Cambridge: Cambridge University Press, 1993. [Online]. Available: https://www.cambridge.org/core/books/computer-user-as-toolsmith/1FFFEE518175E1014E57C7B6B1493863

[18] L. Tauscher and S. Greenberg, "How people revisit web pages: empirical findings and implications for the design of history systems," *International Journal of Human-Computer Studies*, vol. 47, no. 1, pp. 97–137, Jul. 1997. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1071581997901257

[19] A. Montgomery and C. Faloutsos, "Identifying web Browsing Trends and Patterns," *Computer*, vol. 34, no. 7, pp. 94–95, 2001.

[20] R. Kumar and A. Tomkins, "A characterization of online browsing behavior," in *Proceedings of the 19th international conference on World wide web*, ser. WWW '10. New York, NY, USA: Association for Computing Machinery, Apr. 2010, pp. 561–570. [Online]. Available: https://doi.org/10.1145/1772690.1772748

[21] S. Goel, J. Hofman, and M. Sirer, "Who Does What on the web: A Large-Scale Study of Browsing Behavior," *Proceedings of the International AAAI Conference on web and Social Media*, vol. 6, no. 1, pp. 130–137, 2012. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14266

[22] S. K. Tyler, J. Teevan, P. Bailey, S. d. l. Chica, and N. Dandekar, "Large Scale Log Analysis of Individuals' Domain Preferences in web Search," Microsoft Research, Tech. Rep. MSR-TR-2015-048, Jun. 2015. [Online]. Available: https://www.microsoft.com/en-us/research/publication/large-scale-log-analysis-of-individuals-domain-preferences-in-web-search/

[23] H. Müller, J. L. Gove, J. S. webb, and A. Cheang, "Understanding and Comparing Smartphone and Tablet Use: Insights from a Large-Scale Diary Study," in *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction*, ser. OzCHI '15. New York, NY, USA: Association for Computing Machinery, Dec. 2015, pp. 427–436. [Online]. Available: https://dl.acm.org/doi/10.1145/2838739.2838748

[24] Google, "Topics API latest updates | Privacy Sandbox," Nov. 2023. [Online]. Available: https://developers.google.com/privacy-sandbox/relevance/topics/latest

[25] Z. Durumeric, "Cached Chrome Top Million Websites," Feb. 2023. [Online]. Available: https://github.com/zakird/crux-top-lists

[26] V. Le Pochat, T. Van Goethem, S. Tajalizadehkhoob, M. Korczynski, and W. Joosen, "Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation," in *Proceedings 2019 Network and Distributed System Security Symposium*. San Diego, CA: Internet Society, 2019. [Online]. Available: https://www.ndss-symposium.org/wp-content/uploads/2019/02/ndss2019_01B-3_LePochat_paper.pdf

[27] M. Oliveira, J. Yang, D. Griffiths, D. Bonnay, and J. Kulshrestha, "Browsing behavior exposes identities on the Web," Dec. 2023. [Online]. Available: http://arxiv.org/abs/2312.15489

[28] J. Su, A. Shukla, S. Goel, and A. Narayanan, "De-anonymizing Web Browsing Data with Social Networks," in *Proceedings of the 26th International Conference on World Wide Web*. Perth Australia: International World Wide Web Conferences Steering Committee, Apr. 2017, pp. 1261–1269. [Online]. Available: https://dl.acm.org/doi/10.1145/3038912.3052714

[29] W. Aiello, C. Kalmanek, P. McDaniel, S. Sen, O. Spatscheck, and J. Van Der Merwe, "Analysis of Communities of Interest in Data Networks," in *Passive and Active Network Measurement*, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, and C. Dovrolis, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, vol. 3431, pp. 83–96. [Online]. Available: http://link.springer.com/10.1007/978-3-540-31966-5_7

[30] P. McDaniel, S. Sen, O. Spatschek, and J. E. V. D. Merwe, "System and method for tracking individuals on a data network using communities of interest," EP Patent EP1 699 173A1, Sep., 2006. [Online]. Available: https://patents.google.com/patent/EP1699173A1/en

[31] L. Olejnik, C. Castelluccia, and A. Janc, "On the uniqueness of web browsing history patterns," *annals of telecommunications - annales des télécommunications*, vol. 69, pp. 63–74, 2012. [Online]. Available: https://api.semanticscholar.org/CorpusID:14783622

[32] S. Bird, I. Segall, and M. Lopatka, "Replication: Why we still can't browse in peace: On the uniqueness and reidentifiability of web browsing histories," in *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*. Boston, MA, USA: USENIX Association, Aug. 2020, pp. 489–503. [Online]. Available: https://www.usenix.org/conference/soups2020/presentation/bird

[33] Google, "The Privacy Sandbox," 2021. [Online]. Available: https://developer.chrome.com/docs/privacy-sandbox/

Appendix

TABLE 4: Notations and symbols used in this paper.

Notation	Definition	Value
Commit on `Topics` proposal	Commit sha of the `Topics` proposal studied in this paper	*78b13c3* (February 19, 2024)
Taxonomy version	Version of the taxonomy used in this paper	2
Model version	Version of the studied `Topics` classifier in this paper	4
Utility buckets version	Version of the studied `Topics` utility buckets in this paper	1
CrUX version	Version of the CrUX top-list used in this paper	202401 (January 2024)
Tranco version	Version of the Tranco top-list used in this paper	6JZJX (February 4, 2024)
<*browsingTopics()*>	`Topics` API call	*document.browsingTopics()*
n	Number of users	-
p	Probability to pick a random (noisy) topic from taxonomy	0.05
$q = 1 - p$	Probability to pick a real topic from user's top T topics	0.95
T	Number of top topics per epoch	5
τ	Number of topics returned by <*browsingTopics()*>	3 maximum
t_j	Topic j	-
Ω	Number of topics in taxonomy	469 topics (+ *Unknown* topic)