Inferring Users' Demographics and Sensitive Interests Using the Topics API

Athicha Srivirote
srivirote.a@northeastern.edu
Northeastern University
Boston, Massachusetts, USA

Muhammad Abu Bakar Aziz
aziz.muh@northeastern.edu
Northeastern University
Boston, Massachusetts, USA

Jeffrey Gleason
gleason.je@northeastern.edu
Northeastern University
Boston, Massachusetts, USA

Desheng Hu
desheng@ifi.uzh.ch
University of Zurich
Zurich, Switzerland

Christo Wilson
c.wilson@northeastern.edu
Northeastern University
Boston, Massachusetts, USA

Abstract

In 2019, Google introduced the Privacy Sandbox, a collection of APIs designed to facilitate privacy-preserving online advertising. One of the Sandbox APIs, known as Topics, maps users' browsing history to a set of commercially-focused topics and then shares the top topics with advertisers. Prior work has shown that the Topics API makes Chrome users vulnerable to cross-context reidentification attacks. In this work, we investigate whether the Topics API can be abused to implement a different privacy attack: accurate inferences of users' demographics and sensitive interests. To answer this question, we use a real-world dataset of browsing histories—containing over 250,000 unique domains over a span of eight months—to train machine learning models that take topics from the Topics API as input. Of the 19 demographic traits and sensitive interests that we evaluate, we find that all but two show predictive signals. Our findings add to the growing body of evidence that the Topics API is privacy-revealing, not privacy-preserving.

CCS Concepts

• **Security and privacy** → **Privacy protections**; • **Social and professional topics** → *User characteristics*; • **Information systems** → *Online advertising*.

Keywords

Privacy, Google Privacy Sandbox, Topics API

ACM Reference Format:
Athicha Srivirote, Muhammad Abu Bakar Aziz, Jeffrey Gleason, Desheng Hu, and Christo Wilson. 2026. Inferring Users' Demographics and Sensitive Interests Using the Topics API. In *Proceedings of the ACM Web Conference 2026 (WWW '26), April 13–17, 2026, Dubai, United Arab Emirates.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3774904.3792660

1 Introduction

Many browser vendors—such as Apple, Mozilla, and Brave—have restricted or disabled third-party cookies to improve privacy for their users [11, 35]. Google has chosen a different path. In 2019, Google introduced the *Privacy Sandbox*: a suite of APIs in Chrome and Android meant to facilitate privacy-preserving online advertising.[1] The Privacy Sandbox APIs were tailored to common use cases for advertisers, such as remarketing (via the Protected Audiences API) and conversion tracking (via the Attribution Reporting API). Although Google initially claimed that the Privacy Sandbox would replace third-party cookies in Chrome, they have since backtracked [8]. As of October 2025, third-party cookies remain in Chrome and the Privacy Sandbox has been deprecated [9].

Although Google touted the privacy benefits of the Privacy Sandbox APIs, these claims repeatedly failed to stand up to scrutiny. For example, Google asserted that the Privacy Sandbox APIs preserve anonymity by preventing advertisers from reidentifying individual users across contexts (i.e., different websites). However, researchers from Mozilla and academia determined that three Sandbox APIs leaked sufficient information that they could be abused to reidentify users with high accuracy [1, 4, 5, 7, 20, 21, 24, 32, 36, 38].

In this study, we investigate a different privacy attack: whether the Topics API [37] (which we abbreviate as *Topics*) could have been abused to accurately infer users' sensitive attributes. Topics mapped each domain name in users' browsing history to one-or-more commercially-focused *topics of interest* (or, simply, *topics*) drawn from a taxonomy of 469 options (see Table 2 for examples). It then shared each user's top five topics with advertisers on a weekly rotating basis. Since Topics exposed a version of users' browsing history to third parties, it could potentially leak two kinds of sensitive attributes:

(1) **Demographics**: Prior work has observed that users' browsing history is correlated with demographics, such as race and income [4]. It was unknown, however, whether these correlations were maintained after Topics permuted and truncated browsing history.
(2) **Sensitive interests**: The Topics taxonomy did not include sensitive interests (e.g., pornography or health conditions). However, as we show in § 6.1, the process that Chrome used to map domains to topics was not 100% accurate. For example, Topics mapped *xtits.com* (a pornography website) to four topics, including *Movies* and *Celebrities & Entertainment*

WWW '26, Dubai, United Arab Emirates

ACM ISBN 979-8-4007-2307-0/2026/04
https://doi.org/10.1145/3774904.3792660

[1] https://privacysandbox.google.com/

News. Thus, Topics could potentially expose users' sensitive interests to advertisers if Chrome misclassified those domains to seemingly benign, proxy topics.

Prior studies on Topics did not investigate these two privacy threats [1, 5, 20, 21, 36].

In the spirit of Berke and Calacci [4], we present a post-mortem privacy analysis of the Topics API, centered around two research questions:

(1) **RQ1**: Could an adversary have trained machine learning (ML) models that accurately predicted users' demographics and sensitive interests when given topics as input?
(2) **RQ2**: How could changes to the design of Topics affect the accuracy of demographic and sensitive interest predictions?

RQ1 addresses whether Topics was leaking users' private information. **RQ2** is motivated by the fact that Google included mechanisms in Topics that were meant to enhance users' privacy (see § 2). Given this, we seek to understand whether Google could have changed or improved Topics to hinder our attacks.

To answer these questions, we use a data-driven, simulation approach. We rely on a dataset of web browsing histories collected from 782 U. S. residents over eight months in 2020. We input the domains in each participant's browsing history into a simulator that faithfully reproduces the topics generated by the Topics API [5]. Additionally, we use data from FortiGuard to label participants whose browsing history includes sensitive interests [43]. Finally, we train ML models on the topics assigned to participants and assess whether they successfully predict participants' demographic attributes and sensitive interests [2]. Our approach builds on and extends the approaches used by prior work to examine reidentifiability attacks against the Privacy Sandbox [1, 4, 5, 20, 21].

We make the following contributions:

- We demonstrate that of the six demographic traits and 13 sensitive interests that we evaluate, all but two—race and political affiliation—show predictive signals.
- We show that our attacks are robust. Our models show predictive power even when we artificially limit our dataset to 100 participants, limit the advertiser to ten or less calls to the Topics API per epoch, or increase the frequency at which Topics returns random topics to 50% (the default is 5%).

Our findings add to the body of evidence that the Topics API in particular, and the Privacy Sandbox in general, were privacy-revealing, not privacy-preserving.

2 Background

In this study, we focus on the Privacy Sandbox API that was designed to facilitate interest-based ad targeting: Topics [37]. Google proposed Topics in 2022 after its predecessor, FLoC, was deprecated (see § 3). The proposal was rejected by other browser vendors.[3] Nonetheless, Google publicly released Topics in Chrome in Fall 2023.[4] Google announced that most of the Privacy Sandbox APIs, including Topics, would be deprecated starting in October 2025 [9].

[2]https://github.com/athicha/www-2026-google-sandbox-topics-api
[3]See https://github.com/WebKit/standards-positions/issues/111 and https://github.com/mozilla/standards-positions/issues/622.
[4]https://web.archive.org/web/20251124174713/https://www.privacysandbox.com/intl/en_us/open-web/

Parameter	Default Value	Interpretation
e	1 week	Epoch length
t	5	# of top topics chosen per epoch
l	3	# of past epochs to consider, and # of topics returned by *browsingTopics()*
Taxonomy	version 2	Taxonomy of topics; version 2 contains 469 topics
v_o	version 2	Curated override list that maps domains to topics; version 2 contains 47,128 domains
v_b	version 5	BERT model that maps domains to topics
α	0.25	Confidence threshold for the BERT model
p	0.05	Probability that *browsingTopics()* will replace a true topic with a random topic

Table 1: Key parameters to Topics and their default values in Chrome as of early 2025. Unless otherwise specified, we adopt these default values in our Topics simulator.

At a high-level, the goal of Topics was to provide advertisers with each Chrome user's interests, for the purpose of targeting relevant advertisements. Historically, advertisers implemented interest-based ad targeting by tracking users' online behavior and constructing per-user profiles [3, 31, 44]. Topics was designed for a hypothetical world where advertisers no longer track users or construct profiles. Instead, Topics—operating locally within the Chrome browser—performed the tracking and profile construction. By storing browsing history and computing interest profiles locally, Topics aimed to improve privacy for users by exposing less browsing history information to advertisers.

As shown in Figure 1, Topics operated in a series of steps that culminate with advertisers receiving topics from a given user. Table 1 includes the key parameters of Topics and their default values, which we discuss and manipulate in this study.

(1) Topics divided time into e = one week epochs. During each epoch, Topics recorded the URLs visited by the user.
(2) Topics mapped each visited domain name to a topic from a hierarchical taxonomy. As of 2025, Google's version 2 taxonomy included 469 topics (version 1 included 349 topics). Topics used two methods to map a domain to topics:
   (a) Topics tried to look up the topics assigned to the domain in a static mapping table developed by Google known as the *override list*. As of 2025, the version 2 override list included 47,128 domains (version 1 included 9,245).
   (b) If the domain was not present in the override list, then Topics used a pre-trained *BERT classification model* to predict topics. The model's sole input was a domain name and it produced scores for each topic in the range $[0, 1]$. Topics with scores above $\alpha = 0.25$ were retained. Topics used BERT model version 5 since mid-2024.
(3) In a given epoch i, Topics computed the frequency histograms of topics in each of the previous $l = 3$ epochs (i.e., epochs $i-1$, $i-2$, and $i-3$). Topics produced l sets of topics, each of which contained the top $t = 5$ topics from the corresponding epoch. Thus, in any given epoch, there were at most $t \times l = 5 \times 3 = 15$ topics available to advertisers.
(4) Advertisers called a DOM method made available by Topics, *document.browsingTopics()*, that returned an array of strings containing at most l topics, one drawn uniformly at random from each of the top t sets from the previous l epochs.

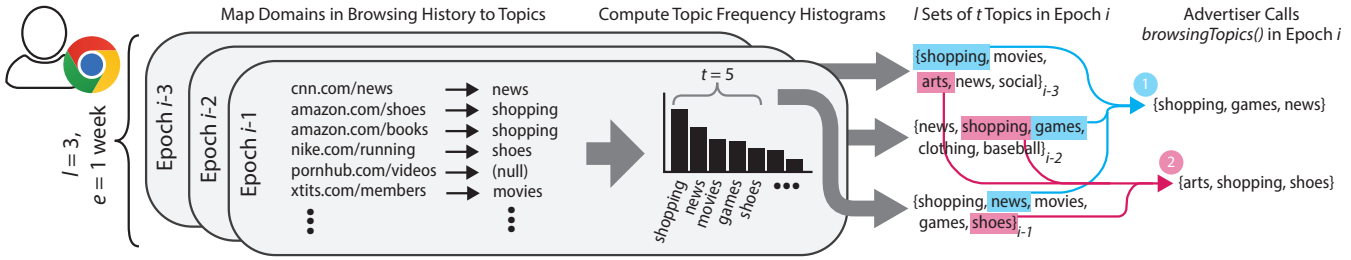

Figure 1: Simplified diagram of Topics in Chrome. In this example, the advertiser calls the DOM *browsingTopics()* from two distinct first-party domains, and thus receives two outputs (one in blue and one in pink).

The Topics API included additional mechanisms that Google claimed improve user privacy [37]. First, there was a $p = 0.05$ chance that *browsingTopics()* would return a topic chosen uniformly at random from the taxonomy, rather than one of the true $t \times l$ topics. Second, when an advertiser invoked *browsingTopics()* on a given first-party domain, the set of l returned topics would remain fixed for that advertiser for the remainder of the epoch. This restriction prevented an advertiser from enumerating all $t \times l$ topics by repeatedly calling the API on a single website. As shown in Figure 1, an advertiser could receive different topics if they invoked *browsingTopics()* on different first-party domains. We discuss how these precautionary mechanisms impact our study in § 4.

Table 2 provides examples of popular topics (in our dataset) from Google's taxonomy. Importantly, the taxonomy did not include sensitive interests, such as pornography, recreational drugs, personal health, gambling, religion, or anything related to race, ethnicity, or sexuality [37]. Google's override list did include domains that fell within these categories, but it mapped them to a special *Unknown* topic or an empty *(null)* topic. An example of this is shown for the domain *pornhub.com* in Figure 1. These two special topics were not included in the calculation of the t topics per epoch.

Google's BERT model also sometimes predicted the *Unknown* topic for domains, but it could make prediction errors, including for domains within sensitive categories. Figure 1 shows a real example of this for the pornography website *xtits.com*. We examine misclassifications by Topics in § 6.1.

3 Related Work

Researchers have examined whether the Privacy Sandbox APIs conformed to Google's privacy claims. For example, Kobayashi et al. [22] found that the Protected Audiences API retained high utility for advertisers. However, two studies found that adversaries could abuse the Protected Audiences API to reidentify individual users across first-party contexts with high accuracy, thus invalidating one of Google's central privacy claims about this API [7, 24].

Researchers from Mozilla analyzed the FLoC API (the predecessor to Topics) and raised concerns that it could be abused to reidentify individual users across contexts [32]. This analysis was confirmed by two studies [4, 38]. Berke and Calacci [4] used data-driven experiments to simulate the functionality of the FloC API, and showed how an adversary could uniquely identify 50% of users based on their FLoC IDs after three weeks of observation, and 95% of users after four weeks.

Pertinent to our study, Berke and Calacci [4] investigated whether FLoC could be abused to infer users' race and income, but did not find statistically significant correlations. We revisit this question for Topics and expand it to cover more demographic traits.

Prior work examined whether Topics could be abused for reidentification attacks. Google produced an analysis purporting to show that Topics could not be abused to reidentify individual users with high accuracy [12], but Mozilla criticized this analysis for adopting a weak threat model and unrealistic statistical assumptions about peoples' web browsing habits [36]. Subsequent research once again confirmed Mozilla's concerns by demonstrating that Topics could be abused to reidentify individual users [1, 5, 20, 21]. These studies used data-driven experiments to simulate the functionality of the Topics API and found that anywhere from 25—57% of users could be uniquely reidentified based on their topics from the Topics API after 15–30 epochs [5, 21]. In their analysis of reidentification risk, Google assumed that the overall distribution of topics over the online population would be normally distributed. However, Beugin and McDaniel [5] found that, under naturalistic browsing patterns, the topics returned by the Topics API were not uniformly distributed. For example, they found that the most popular topic was returned on 18% of websites (similar to our results, see Table 2), while 196 of the topics were never returned.

Although Google added noise to Topics to try and enhance privacy, Jha et al. [21] determined that a simple denoising algorithm was sufficient to filter out the random topics inserted by Google's algorithm, thus obviating its privacy benefits.

We build on and extend this prior work. Although prior work investigated whether Topics could be abused to reidentify individual web users, no studies examined whether it could be abused to infer users' demographics or sensitive interests. Like prior work, we aim to answer this question through data-driven simulations.

4 Threat Model

In this study, we simulate an advertiser whose goal is to infer users' demographics and sensitive interests from the topics produced by the Topics API so that they may target ads to these personal characteristics [2, 34]. For the purposes of model training, we assume that the advertiser has access to a dataset that contains the ground-truth demographics and browsing histories of a sample population. At inference time, we assume that the advertiser has the same capabilities as any other third-party on the web: they may observe visitors to websites and call the Topics API.

Topic	Top Three Example Domains	#	%
/Internet & Telecom	mail.google.com, google.com, outlook.live.com	10.5 M	17.86
/Online Communities	facebook.com, youtube.com, twitter.com	5.7 M	9.71
/Arts & Entertainment	youtube.com, yahoo.com, chess.com	4.5 M	7.59
/Online Communities/Social Networks	facebook.com, twitter.com, instagram.com	3.5 M	5.92
/Computers & Electronics/Software	google.com, docs.google.com, translate.google.com	2.9 M	4.96

Table 2: Top five topics assigned to our participants and the top three domains mapped to each topic. We present the total number of URLs that were mapped to each topic and the percentage of all topics that each accounts for. Note that a single domain can be mapped to multiple topics (e.g., *google.com*).

Topics is deterministic (see § 2), so given knowledge of browsing history, the advertiser can compute the topic distribution for each member of the sample population (as prior work has done [5] and we will do in § 5.2). Some advertisers—e.g., Meta—already possess this kind of dataset. Alternatively, an advertiser could purchase the requisite data from a data broker [44] or from a panel provider (which is the tactic we utilize, see § 5.1). We examine the relationship between training dataset size and prediction accuracy in § 6.2.2.

We assume that Chrome implements Topics correctly, including the privacy-enhancing features stated in § 2. We examine the relationship between these privacy-enhancing features and prediction accuracy in § 6.3.

5 Methods

We now describe the implementation of our study, including our dataset of browsing histories, our approach to simulating the Topics API, and our approach to training and evaluating ML models.

5.1 Dataset

We worked with the survey company YouGov to recruit a panel of U. S. residents to take a survey and optionally install a browser extension that was compatible with Chrome and Firefox. The survey collected basic demographic information and the extension collected browsing history. We fielded our data collection from May to December 2020. In total, 782 participants consented, completed the survey, and installed the extension. However, we filtered out the 11 participants with less than seven days of browsing history. After filtering, our dataset includes 37,575,287 total URLs over 258,647 unique domains. See § 6.1 for analysis of participant behavior, and § 8.2 for ethical considerations in our data collection. This dataset has been utilized by several prior works [10, 14–17, 33].

5.1.1 Demographics. Table 3 presents the self-reported demographics of our participants in comparison to the 2020 US Census [39–42]. We observe that our sample under-represents racial minorities, people with no college education, high-income earners, people age 18–34, and Republicans. Our sample over-represents people with post-graduate education and people age 55–64. We revisit this limitation of our study in § 8.1.

As shown in Table 3, when we analyze participants by income, age, and education, we group participants into three ranges. For political affiliation, we focus on the three largest cohorts in our sample: independents, Republicans, and Democrats. For race/ethnicity, we also focus on the four largest cohorts: White, Black, Asian, and Hispanic. Lastly, we analyze participants by binary gender.

	Sample		Census
	N	%	%
Total	771		
Gender			
Male	368	0.477	0.495
Female	403	0.523	0.505
Race			
White	617	0.800	0.759
Black	59	0.077	0.135
Asian	17	0.022	0.061
Hispanic	45	0.058	
Education			
No college	77	0.010	0.376
Some college	474	0.502	0.497
Post-grad	220	0.285	0.127
Household Income			
Less than $50,000	275	0.357	0.378
$50,000–$99,999	266	0.345	0.286
$100,000 or more	181	0.235	0.336
Age			
18–34 years	112	0.145	0.294
35–44 years	277	0.359	0.326
55–64 years	382	0.495	0.380
Party Identification			
Democrat	433	0.561	
Republican	94	0.122	
Independent	193	0.250	

Table 3: Self-reported demographic distributions of our participants, compared to the 2020 US Census.

Category	%
Health and Wellness	0.962
Political Organizations	0.811
Medicine	0.654
Global Religion	0.643
Other Adult Materials	0.503
Pornography	0.472
Alcohol	0.459
Gambling	0.319
Weapons (Sales)	0.239
Marijuana	0.187
Folklore	0.175
Tobacco	0.169
Abortion	0.071

Table 4: Percentage of participants who visited at least one domain in each sensitive category.

5.1.2 Sensitive Interests. To identify participants with sensitive interests, we used data from FortiGuard to classify the domains that participants had visited. Prior work has shown that FortiGuard offers highly accurate categorical labels for domains [43]. In total, FortiGuard includes 90 categories. We manually examined them all and identified the 13 categories shown in Table 4 to be sensitive. because they concern vices, heavily regulated (and potentially illegal) items, deeply personal issues, and personal beliefs.[5]

For each sensitive category, we assigned a binary label to each participant in our dataset indicating whether they did (positive) or did not (negative) visit at least one domain in that category.

5.2 Topics API Simulator

To generate the topics that would have been assigned to our participants by Topics, we built a Topics API simulator. Our simulator takes a participant's browsing history (i.e., URLs and timestamps) as input and outputs their top topics grouped by epoch, calculated using the process described in Google's publicly available documentation[6] and in § 2. We use source code implementing key routines developed by Beugin and McDaniel [5], who validated that their implementation produced topics that perfectly matched Chrome's implementation. Unless otherwise specified, we adopted Chrome's default parameters (see Table 1) in our simulator. To answer **RQ2**, we experiment with non-default parameters in § 6.2.2.

[5]Fortiguard describes each category here: https://www.fortiguard.com/webfilter/categories. Fortiguard describes the "other adult materials" category as "Mature content websites (18+ years and over) that feature or promote sexuality, [. . .] without the intent to sexually arouse."

[6]https://developers.google.com/privacy-sandbox/private-advertising/topics/web

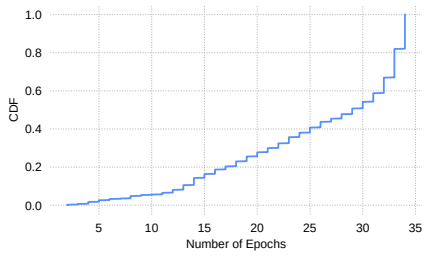

Figure 2: CDF of the number of one week epochs with browsing history data per participant.

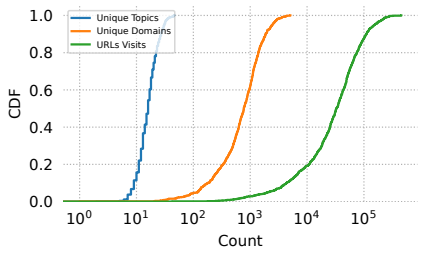

Figure 3: CDF of the unique topics assigned, unique domains visited, and total URLs visited by each participant.

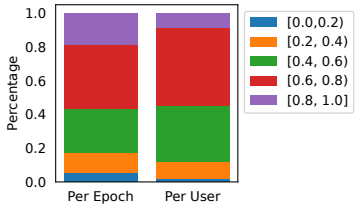

Figure 4: Set overlap between topics in adjacent epochs, per participant. The left bar presents all Jaccard indices. The right bar presents the median Jaccard index per participant.

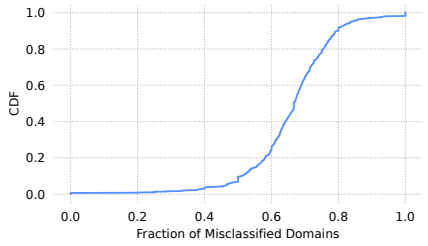

Figure 5: CDF of the fraction of misclassified unique sensitive domains visited by each participant.

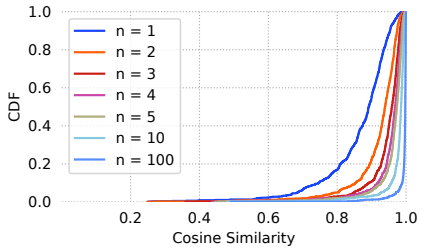

Figure 6: CDF of the cosine similarity between ground truth topic vectors and topic vectors accumulated over n *browsingTopics()* calls, per participant.

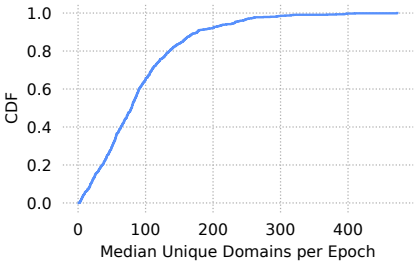

Figure 7: CDF of the median number of unique domains visited per epoch by each participant.

5.3 Prediction and Evaluation

To answer our research questions, we evaluate whether an adversary can infer participants' demographics and sensitive interests from the outputs of Topics.

5.3.1 Inference Setting. The prediction task is structured around two types of input data. During the training phase, the advertiser can map browsing history to topic distributions because both the override list and the BERT model behind Topics are publicly available. During the testing phase, the advertiser must query Topics to obtain a person's topic distribution as shown in Figure 1. This asymmetry makes the task an instance of dataset shift [29].

The amount of information that Topics reveals depends on two parameters: the number of top topics (t) and the probability of returning a random topic (p). We represent advertiser strength as the number of times they can call Topics (n), which directly maps to the number of unique first-party domains that include the advertiser's Javascript. We perform ablation studies that vary the amount of training data available to advertisers (operationalized as the number of participants, m), the amount of information revealed by Topics (t and p), the version of Topics in use (v_o and v_b), and the advertiser's footprint on the web (n).

5.3.2 Feature Construction. We represent each participant's browsing history as a distribution over visited domains. From this, we generate the top t topics per epoch using the Topics API Simulator and compute a matrix of topics for each epoch. We then convert each matrix into a per-user vector by summing the topics in each column and normalizing the frequency counts by the number of epochs. We represent each sensitive attribute as a binary indicator. For categorical traits, we use one-vs-rest coding (e.g., income is split into three indicators: $< \$50K$, \$50K–\$99K, and $\geq \$100K$). We split participants into train (80%) and test (20%) sets using a single stratified random split.

5.3.3 Model Architecture. In our experiments, we train Random Forest (RF) classifiers. We tested a variety of classical ML models and found that the choice had little impact on performance. RFs make sense given our tabular input data with fewer than 1,000 samples and 500 features. For each classifier, we tuned hyperparameters using grid search with 10-fold cross-validation.

5.3.4 Evaluation. We evaluate predictions using the area under the ROC curve (AUC-ROC), which is appropriate for binary classification and robust to class imbalance (see Table 3). A score of 0.5 indicates random guessing, while a score of 1.0 indicates perfect ranking of positive and negative instances. We quantify uncertainty in our main results using 1,000 bootstrap resamples over the test set [30]. For interpretability, we compute SHapley Additive exPlanations (SHAP) [25], which attribute predictive performance to individual topics.

6 Results

We now present the results of our analysis and answer our RQs.

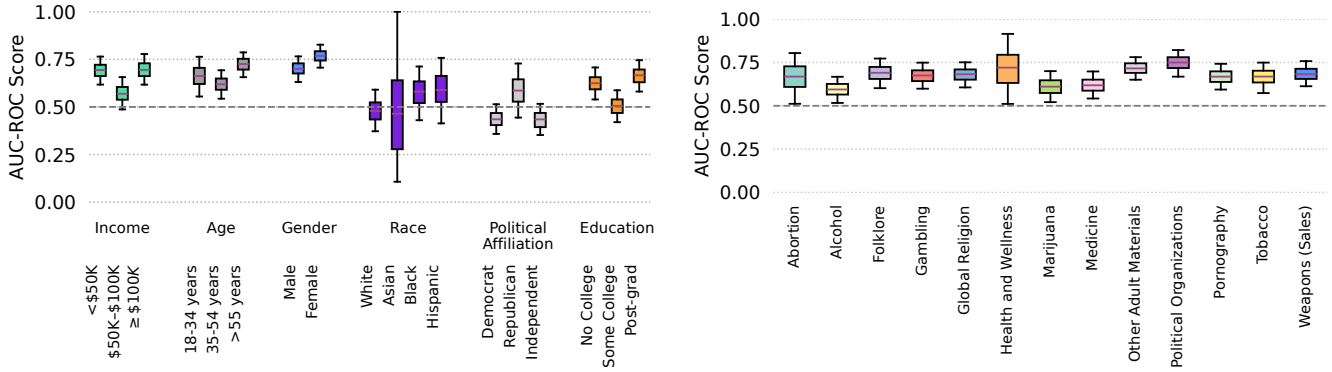

Figure 8: AUC-ROC scores for RF models trained to predict participants' demographics and sensitive interests when given topics as input. All parameters for Topics were set to their default values and $n = \infty$. Boxes show 5, 25, 50, 75, and 95th percentile AUC-ROC scores for each model after 1,000 rounds of bootstrapping. Scores above 0.5 indicate strong predictive performance.

6.1 Participant Behavior

Before delving into our RQs, we first examine the overall behavior of participants in our study. Figure 2 presents an empirical Cumulative Distribution Function (CDF) of the number of one week epochs of data that we collected from each participant. Although we collected data for roughly 34 weeks, not all participants were active for the entire duration. We observe that the median participant provided 29 epochs of data, and the vast majority contributed sufficient data to simulate their topics for many epochs.

Figure 3 presents the CDF of total URLs visited and unique domains visited per participant. While the median participant visited 32,921 URLs (or 308 per day on average), the median participant only visited 772 unique domains. This is expected, as people repeatedly visit popular domains.

Figure 3 also presents the CDF of unique topics assigned to each participant over time by our simulator (only considering the top five topics per epoch). The median participant was assigned 16 topics, while the maximum topics assigned to one participant was 46. These quantities are an order of magnitude lower than the number of unique domains visited by participants, but this is explained by the fact that many domains map to the same topics (see Table 2 for examples).

Overall we observed 443 topics being returned by the Topics API throughout the course of our simulations, out of the 469 available in the taxonomy. This is a similar ratio to the 342 out of 349 available topics observed by Beugin and McDaniel [5].

Prior work has observed that people tend to have stable web browsing patterns, which leads to stable topics over time [6]. To examine this phenomenon, we compute the Jaccard index between the five topics assigned to each participant in adjacent epochs. Figure 4 presents stacked bars showing the distribution of all Jaccard indices on the left, and the median Jaccard index per participant on the right. In both cases, ~60% of adjacent epochs include four or five topics that match. This is greater stability than Beugin and McDaniel [6] observed: they reported that ~21% of the users in their dataset had four or five stable topics epoch-to-epoch. This discrepancy is potentially explained by the fact that participants in our study contributed data for 34 weeks, while the participants in the Beugin and McDaniel [6] study contributed data for five weeks.

Finally, we investigate cases where Topics misclassified participants' browsing history. We define misclassifications as cases where domains in a sensitive interest category (according to Fortiguard, see § 5.1.2) are not mapped to the *Unknown/null* topic by Topics. Figure 5 presents the distribution of fraction of misclassified sensitive domains per participant. We see that misclassifications are the norm: for the median participant, 66.7% of the sensitive domains they visit are misclassified, with the worst case being 100%. These results motivate our investigation of whether Topics is leaking sufficient information to enable prediction of sensitive interests.

6.2 RQ1: Prediction Accuracy

Next, we assess **RQ1**: can an advertiser train ML models that accurately predict the sensitive attributes of participants based on their topics? Figure 8 shows the distribution of AUC-ROC scores for each model over 1000 bootstrap resamples of the test set. (Table 5 in the Appendix presents the precision, recall, and F1 scores for these models.) We configure Topics to its default parameters (see Table 1). In this experiment we set $n = \infty$, i.e., the advertiser may call Topics an unlimited number of times per participant each epoch (see § 5.3). This represents a best-case scenario for the advertiser, as they can learn the true topics for each participant and eliminate random topics [21]. We investigate weaker attackers in the § 6.3.

Most of our models exhibit at least some predictive power. Nine demographic attribute models and twelve sensitive interest models have median AUC-ROC scores above 0.6. Three demographic attribute models (≥ 55 years old Male, and Female) and three sensitive interest models (Health & Wellness, Other Adult Materials, and Political Organizations) have median scores above 0.7. Thus, Topics does leak sufficient information for an advertiser to predict many demographic traits and sensitive attributes effectively.

The seven models predicting race and political affiliation perform poorly. Race is the most unbalanced class in our dataset (see Table 3) which explains the inconsistent performance of these models.

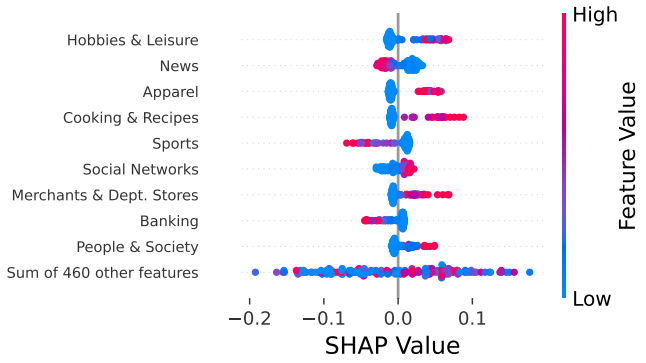

Figure 9: SHAP values for the demographic attribute model that predicts whether a participant is female.

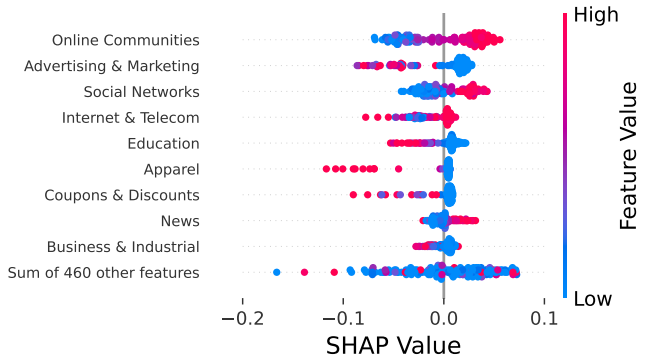

Figure 10: SHAP values for the sensitive interest model that predicts interest in Political Organizations.

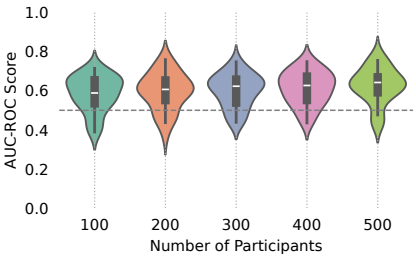

Figure 11: Distribution of AUC-ROC scores for all 31 of our models as we vary the number of participants m. $n = \infty$.

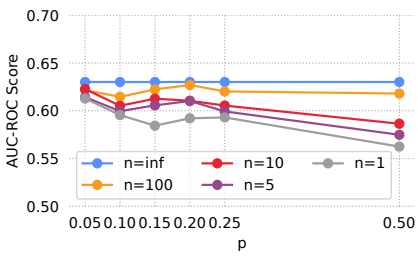

Figure 12: Average AUC-ROC scores across our 31 models as we vary p and n.

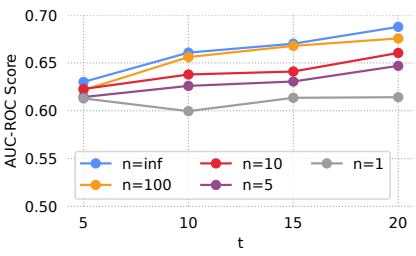

Figure 13: Average AUC-ROC scores across our 31 models as we vary t and n.

6.2.1 Feature Importance. To explain what patterns our models exploit, we present SHAP values for the best demographic attribute model (Female) and the best sensitive interest model (Political Organizations). Figure 9 and Figure 10 show the ten most important features (i.e., topics) for each model and summarize the distribution of SHAP values over instances in the test set. Red dots represent high feature values (e.g., high frequency of the Cooking & Recipe topic) and the x-axis represents SHAP values (i.e., the importance of a topic to the model's prediction). We also show SHAP values for the second best performing models in each category (≥ 55 years old and Other Adult Materials) in the Appendix.

High frequencies of the Hobbies & Leisure, Apparel, Cooking & Recipes, Mass Merchants & Department Stores, and People & Society topics increase the model's expectation that a participant is female, while high frequencies of the News, Sports, and Banking topics reduces its expectation. High frequencies of the Online Communities, Social Networks, and News topics increase the model's expectation that a participant is interested in political organizations, while frequencies of the Advertising & Marketing, Education, Apparel, and Coupons & Discounts topics reduces its expectation. These explanations offer face validity to the correlations that our models leverage. They also demonstrate that some of the proxies for demographic traits and sensitive attributes in the Topics taxonomy are stereotypical. We present additional examples in § A.1.

6.2.2 Impact of Training Data Size. The baseline models in Figure 8 were trained on $m = 616$ participants (i.e., 80% of the dataset). To investigate the impact of training data size on model performance, we varied the number of participants used for training, while holding the test dataset fixed. Figure 11 shows the distribution of AUC-ROC scores across our 31 models as we vary m from 100 to 500. As expected, the median AUC-ROC score increases from 0.59 when $m = 100$ to 0.64 when $m = 500$. Models trained on more participants produce better inferences about demographic attributes and sensitive interests. However, even with access to only $m = 100$ participants, almost half of our models have AUC-ROC scores greater than 0.6. Thus, even advertisers who can only acquire datasets much smaller than ours can still make useful inferences about a number of sensitive attributes. On the other hand, advertisers who can access data on more than $m = 616$ participants can likely produce even stronger inferences about sensitive attributes.

6.3 RQ2: Impact of Changes to Topics

Next, we address **RQ2**: how do changes to Topics impact the accuracy of predictions? We investigate three variables—p, t, and n—in this section and two more—v_o and v_b—in § A.2 and § A.3.

p is the parameter of Topics that controls how frequently the API returns random topics to advertisers, and thus it is Google's primary mechanism for trading off privacy versus utility. Figure 12 examines the accuracy of our models as p varies from 0.05 (its default value)

to 0.5. We also vary $n = [1, 5, 10, 100, \infty]$ to explore the interplay between the advertiser's power and Google's countermeasure. Each point in Figure 12 is the average AUC-ROC score computed across our 31 models. Aside from p, we leave all other Topics parameters at their default values.

Figure 12 reveals that the average performance of our models does tend to decrease as p increases. However, the decrease in average AUC-ROC score is not dramatic—0.06 in the worst case when comparing $p = 0.05$ to $p = 0.5$—and decreases are constrained to instances where the advertiser has limited power, i.e., $n \leq 10$. These results demonstrate that increasing p, even to an impractical amount, is insufficient to prevent Topics from leaking information about sensitive attributes.

Figure 13 examines the relationship between t (the number of top topics selected each epoch), n, and the average AUC-ROC scores of our models. We observe that, were Google to increase t, this would increase average predictive performance. This is especially true as n increases, as this affords the advertiser more opportunity to collect these additional topics from participants. These result highlight that Topics could potentially leak more information about users' sensitive attributes in the future if, for example, Google decided to increase the utility of Topics to advertisers by increasing t.

Figure 12 and Figure 13 both show that the average AUC-ROC score of our models increase rapidly as n grows, but that by $n = 100$ the average AUC-ROC score approaches what our models can achieve under the idealized conditions of $n = \infty$. To explain this result, we plot Figure 6, which shows the distribution of cosine similarities between the ground truth topic vectors and the computed topic vectors after n calls to *browsingTopics()*, per participant. We set $p = 0.05$ in this experiment. We observe that an advertiser is able to learn the vast majority of true topics for participants in relatively few calls to the Topics API. Figure 6 shows that median cosine similarity increases rapidly, from 0.89 when $n = 1$, to 0.97 when $n = 5$, to 0.99 when $n = 10$. To put this into further perspective, Figure 7 presents the median number of unique domains visited per epoch per participant. 95% of participants visited at least 10 unique domains per epoch, meaning that it is very feasible for an advertiser to achieve $n \geq 10$ in practice.

Taken together, our observations in this section and in § A.2 and § A.3 demonstrate that our findings about privacy leakage from the Topics API are robust under a variety of conditions.

7 Conclusion

In this study, we use browsing history data from 782 US residents to examine whether Chrome's Topics API leaked six demographic traits and 13 sensitive interests. We use a high-fidelity simulator to generate the topics that would have been assigned to these participants. Our study is the first to show that ML models trained on topics data can predict unseen demographics and sensitive interests (**RQ1**). Only race and political affiliation were unpredictable. We examined the features learned by our models and found that they are identifying browsing behaviors that are stereotypically correlated with specific demographic traits and interests. Additionally, we showed that the predictive performance of our models are robust even when parameters of the Topics algorithm are changed (**RQ2**).

8 Discussion

How "bad" are our findings, in a normative sense? On one hand, Google never explicitly promised that Topics would not leak users' demographics or sensitive interests (they only discussed reidentification attacks [12]). On the other hand, Google was clearly aware that protecting the privacy of sensitive attributes was important: Google intentionally omitted these items from the Topics taxonomy,[7] and Google intentionally—albeit imperfectly, see Figure 5—mapped many sensitive domains to the *Unknown* topic. Thus, while we cannot claim to have caught Google breaking an explicit promise to users, our findings do complicate Google's claim that Privacy Sandbox was a privacy-enhancing technology.

Given our results and those from prior work [1, 4, 5, 7, 20, 21, 24, 32, 36, 38], it is clear that the Privacy Sandbox APIs did not sufficiently protect users' privacy, and thus we applaud Google's decision to deprecate them. That said, we do support the concept of building privacy-preserving advertising APIs into browsers [45], provided that (1) those APIs offer strong, rigorously verified privacy guarantees and (2) are paired with the deprecation of privacy-violating functionality, such as third party cookies. As the the developer of the world's most popular browser—Chrome—and the preeminent adtech firm on the internet, Google will need to be involved in these developments. However, we argue that Google should not lead these developments, given their dominant position in multiple key markets [26, 27] and their inherent conflict of interest, i.e., as a platform developer and an advertiser.

8.1 Limitations and Future Work

Our study relies on a browsing history dataset gathered from 782 US residents. It is possible that our findings may not generalize to larger populations, or populations outside the US. While our model evaluation approach accounts for class imbalances, it is possible that our results might change if our sample was perfectly representative of the US population. Future work could improve on this study by utilizing browsing history data from a larger and more representative user population.

We do not estimate the monetary value of our models' predictions to advertisers, and thus we cannot concretely assess advertisers' incentive to implement privacy attacks against Topics. Assessing the value of data in advertising markets—especially uncertain data—is an under-explored area of research [23].

8.2 Ethical Principles

This study utilizes sensitive data collected from participants that cannot be publicly released. Our protocol was approved under Northeastern IRB #20-03-04. Participants were informed about our data collection and required to consent before we recorded any data. Participants were compensated and were free to exit the study at any time, including asking us to delete all collected data. We did not receive any requests to opt-out or delete data.

[7] In contrast, Google's other targeted advertising taxonomies do incorporate demographics and sensitive interests [3, 18, 19].

References

[1] Mário S. Alvim, Natasha Fernandes, Annabelle McIver, and Gabriel H. Nunes. 2024. The Privacy-Utility Trade-off in the Topics API. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*.

[2] Athanasios Andreou, Marcio Silva, Fabricio Benevenuto, Oana Goga, Patrick Loiseau, and Alan Mislove. 2019. Measuring the Facebook Advertising Ecosystem. In *Proceedings of the USENIX Symposium on Networked Systems Design and Implementation*.

[3] Muhammad Ahmad Bashir, Umar Farooq, Maryam Shahid, Muhammad Fareed Zaffar, and Christo Wilson. 2019. Quantity vs. Quality: Evaluating User Interest Profiles Using Ad Preference Managers. In *Proceedings of the Network and Distributed System Security Symposium*.

[4] Alex Berke and Dana Calacci. 2022. Privacy Limitations of Interest-based Advertising on The Web: A Post-mortem Empirical Analysis of Google's FLoC. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*.

[5] Yohan Beugin and Patrick McDaniel. 2024. Interest-disclosing Mechanisms for Advertising are Privacy-Exposing (not Preserving). *Proceedings on Privacy Enhancing Technologies* 2024, 1 (2024), 41–57.

[6] Yohan Beugin and Patrick McDaniel. 2024. A Public and Reproducible Assessment of the Topics API on Real Data. In *Proceedings of the IEEE Security and Privacy Workshops*.

[7] Giuseppe Calderonio, Mir Masood Ali, and Jason Polakis. 2024. Fledging Will Continue Until Privacy Improves: Empirical Analysis of Google's Privacy-Preserving Targeted Advertising. In *Proceedings of the USENIX Security Symposium*.

[8] Anthony Chavez. 2024. A new path for Privacy Sandbox on the web. The Privacy Sandbox. https://privacysandbox.google.com/blog/privacy-sandbox-update.

[9] Anthony Chavez. 2025. Update on Plans for Privacy Sandbox Technologies. The Privacy Sandbox. https://privacysandbox.google.com/blog/update-on-plans-for-privacy-sandbox-technologies.

[10] Annie Y. Chen, Brendan Nyhan, Jason Reifler, Ronald E. Robertson, and Christo Wilson. 2023. Subscriptions and external links help drive resentful users to alternative and extremist YouTube videos. *Science Advances* 9, 35 (Aug. 2023).

[11] Clifford Colby and Rae Hodge. 2021. This privacy-focused browser stops websites tracking you even better than Chrome does. CNet. https://www.cnet.com/tech/mobile/a-privacy-focused-browser-that-stops-websites-from-tracking-you-even-better-than-chrome-does/.

[12] Alessandro Epasto, Andres Munoz Medina, Christina Ilvento, and Josh Karlin. 2022. Measures of cross-site re-identification risk: An analysis of the Topics API Proposal. Private Advertising Technology Community Group Individual Draft Space. https://github.com/patcg-individual-drafts/topics/blob/main/topics_analysis.pdf.

[13] Michelle Faverio and Olivia Sidoti. 2024. Teens, Social Media and Technology 2024. Pew Research Center. https://www.pewresearch.org/internet/2024/12/12/teens-social-media-and-technology-2024/.

[14] Jeffrey Gleason, Desheng Hu, Ronald E. Robertson, and Christo Wilson. 2023. Google the Gatekeeper: How Search Components Affect Clicks and Attention. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*.

[15] Jeffrey Gleason, Alice Koeninger, Desheng Hu, Jessica Teurn, Yakov Bart, Samsun Knight, Ronald E. Robertson, and Christo Wilson. 2024. Search Engine Revenue from Navigational and Brand Advertising. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*.

[16] Desheng Hu, Jeffrey Gleason, Muhammad Abu Bakar Aziz, Nikolas Guggenberger, Ronald E. Robertson, and Christo Wilson. 2024. Market or Markets? Investigating Google Search's Market Shares Under Vertical Segmentation. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*.

[17] Desheng Hu, Ronald E. Robertson, Aniko Hannak, and Christo Wilson. 2024. U. S. Users' Exposure to YouTube Videos On- and Off-platform. In *Proceedings of the ACM Conference on Web Science*.

[18] IAB Tech Lab 2024. Audience Taxonomy. https://iabtechlab.com/standards/audience-taxonomy/.

[19] IAB Tech Lab 2024. Content Taxonomy. https://iabtechlab.com/standards/content-taxonomy/.

[20] Nikhil Jha, Martino Trevisan, Emilio Leonardi, and Marco Mellia. 2023. On the Robustness of Topics API to a Re-Identification Attack. *Proceedings on Privacy Enhancing Technologies* 2023, 4 (2023), 66–78.

[21] Nikhil Jha, Martino Trevisan, Emilio Leonardi, and Marco Mellia. 2024. Re-Identification Attacks against the Topics API. *ACM Trans. Web* 18, 3 (Aug. 2024).

[22] Shunto Kobayashi, Garrett Johnson, and Zhengrong Gu. 2024. Privacy-Enhanced versus Traditional Retargeting: Ad Effectiveness in an Industry-Wide Field Experiment. SSRN. https://ssrn.com/abstract=4972368.

[23] Xiao-Bai Li, Xiaoping Liu, and Luvai Motiwalla. 2021. Valuing Personal Data with Privacy Consideration. *Decision Sciences* 52, 2 (2021), 393–426.

[24] Minjun Long and David Evans. 2024. Evaluating Google's Protected Audience Protocol. *Proceedings on Privacy Enhancing Technologies* 2024, 4 (2024), 892–906.

[25] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*.

[26] David McCabe. 2025. Fate of Google's Ad Tech Monopoly Is Now in a Judge's Hands. New York Times. https://www.nytimes.com/2025/11/21/technology/google-ad-tech-closing.html.

[27] David McCabe. 2025. Fate of Google's Search Monopoly Is Now in a Judge's Hands. New York Times. https://www.nytimes.com/2025/05/30/technology/google-search-antitrust-chrome.html.

[28] Pew 2024. Social Media Fact Sheet. Pew Research Center. https://www.pewresearch.org/internet/fact-sheet/social-media/.

[29] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. 2022. *Dataset shift in machine learning*. Mit Press.

[30] Sebastian Raschka. 2022. Creating Confidence Intervals for Machine Learning Classifiers. Blog post. https://sebastianraschka.com/blog/2022/confidence-intervals-for-ml.html Accessed on 2025-09-24.

[31] Nathan Reitinger, Bruce Wen, Michelle L. Mazurek, and Blase Ur. 2023. Analysis of Google Ads Settings Over Time: Updated, Individualized, Accurate, and Filtered. In *Proceedings of the Workshop on Privacy in the Electronic Society*.

[32] Eric Rescorla and Martin Thomson. 2021. Technical Comments on FLoC Privacy. Mozilla Technical Report. https://mozilla.github.io/ppa-docs/floc_report.pdf.

[33] Ronald E. Robertson, Jon Green, Damian J. Ruck, Katherine Ognyanova, Christo Wilson, and David Lazer. 2023. Users choose to engage with more partisan news than they are exposed to on Google Search. *Nature* 618 (May 2023).

[34] Till Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Nunes Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna P. Gummadi, Patrick Loiseau, and Alan Mislove. 2018. On the Potential for Discrimination in Online Targeted Advertising. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*.

[35] Nick Statt. 2020. Apple updates Safari's anti-tracking tech with full third-party cookie blocking. The Verge. https://www.theverge.com/2020/3/24/21192830/apple-safari-intelligent-tracking-privacy-full-third-party-cookie-blocking.

[36] Martin Thomson. 2023. A Privacy Analysis of Google's Topics Proposal. Mozilla Technical Report. https://mozilla.github.io/ppa-docs/topics.pdf.

[37] Topics 2023. Topics API for Web. Google Privacy Sandbox Developer Documentation. https://developers.google.com/privacy-sandbox/private-advertising/topics/.

[38] Florian Turati, Karel Kubicek, Carlos Cotrini, and David Basin. 2023. Locality-Sensitive Hashing Does Not Guarantee Privacy! Attacks on Google's FLoC and the MinHash Hierarchy System. *Proceedings on Privacy Enhancing Technologies* 2023, 4 (2023), 117–131.

[39] U.S. Census Bureau. 2020. Educational Attainment in the United States: 2020. https://www.census.gov/data/tables/2020/demo/educational-attainment/cps-detailed-tables.html. All Races file: table 1-1.

[40] U.S. Census Bureau 2021. HINC-01: Selected Characteristics of Households by Total Money Income in 2020. https://www.census.gov/data/tables/time-series/demo/income-poverty/cps-hinc/hinc-01.2020.html. Current Population Survey (CPS) 2021 Annual Social and Economic Supplement, Table HINC-01.

[41] U.S. Census Bureau 2024. Annual Estimates of the Resident Population by Sex, Race, and Hispanic Origin for the United States: April 1, 2020 to July 1, 2023 (NC-EST2023-SR11H). https://www.census.gov/data/tables/time-series/demo/popest/2020s-national-detail.html.

[42] U.S. Census Bureau 2024. Monthly Population Estimates by Age, Sex, Race and Hispanic Origin for the United States: 7/1/2020 to 12/1/2020. https://www2.census.gov/programs-surveys/popest/datasets/2020-2023/national/asrh/nc-est2023-alldata-r-file02.csv.

[43] Pelayo Vallina, Victor Le Pochat, Álvaro Feal, Marius Paraschiv, Julien Gamba, Tim Burke, Oliver Hohlfeld, Juan Tapiador, and Narseo Vallina-Rodriguez. 2020. Mis-shapes, mistakes, misfits: An analysis of domain classification services. In *Proceedings of the ACM Internet Measurement Conference*.

[44] Giridhari Venkatadri, Piotr Sapiezynski, Elissa M. Redmiles, Alan Mislove, Oana Goga, Michelle Mazurek, and Krishna P. Gummadi. 2019. Auditing Offline Data Brokers via Facebook's Advertising Platform. In *Proceedings of the World Wide Web Conference*.

[45] Christo Wilson. 2023. In Support of Standards for Digital Advertising. *Harvard Journal of Law & Technology* 37, 3 (2023), 1063–1085.

A Appendix

A.1 Additional SHAP Explanations

Figure 14 presents the SHAP explanations for our model that predicted likelihood of being 55 years of age or older. High frequencies of the News, Social Networks, and Law & Government topics increase the model's expectation that a participant is 55 years or older. Note that the top two most visited social network sites in our dataset are Facebook (~2M visits) and Twitter (~702K visits), which have an older user base than the third most visited social network,

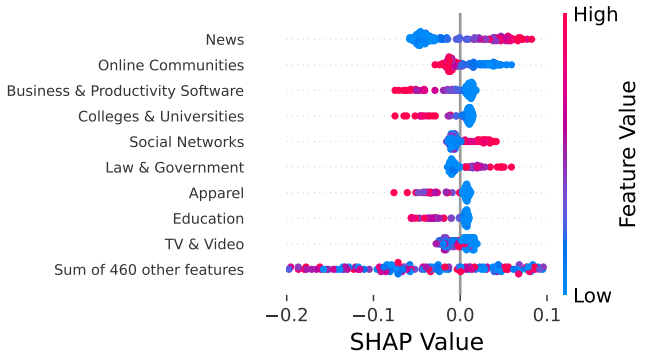

Figure 14: SHAP values for the demographic attribute model that predicts whether a participant is 55 years or older. Red dots represent high feature values (e.g., high frequency of the News topic) and the x-axis represents SHAP values (i.e., the importance of a topic to the model's prediction)

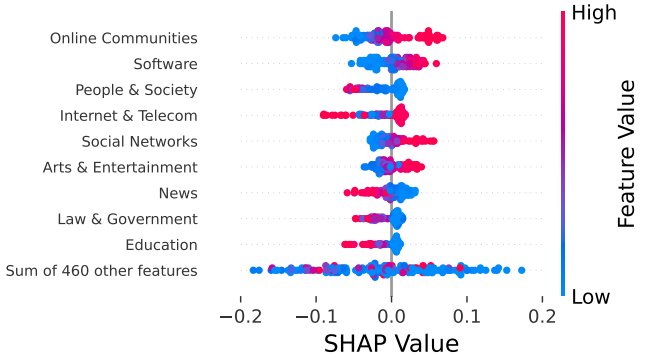

Figure 15: SHAP values for the sensitive interest model that predicts interest in Adult Materials. Red dots represent high feature values (e.g., high frequency of the Online Communities topic) and the x-axis represents SHAP values (i.e., the importance of a topic to the model's prediction)

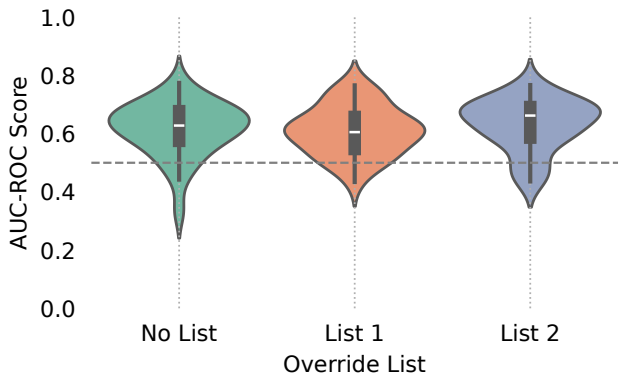

Figure 16: Distribution of AUC-ROC scores for all of our 31 models as we vary what (if any) version of the override list is used in Topics.

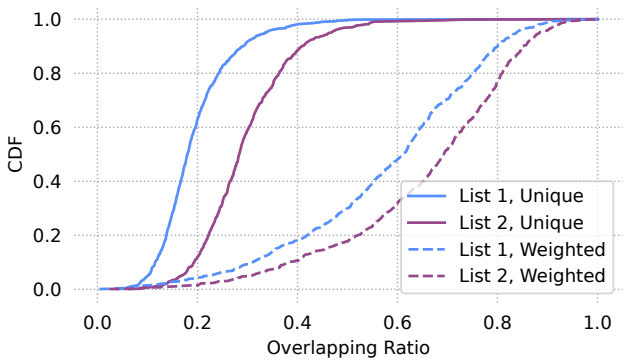

Figure 17: CDF of overlap between visited domains in the override list and all domains visited, per participant. Solid lines are calculated using set overlap considering unique domains. Dashed lines are weighted to account for the visit frequency of each domain, per participant.

Instagram (~118K visits) [13, 28]. In contrast, high frequencies of the Online Communities, Business & Productivity Software, and Colleges & Universities topics reduces the model's expectation that a participant is 55 years or older.

Figure 15 presents the SHAP explanations for our model that predicted interest in the Other Adult Materials sensitive interest. High frequencies of the Online Communities, Software, Social Networks, and Arts & Entertainment topics increase the model's expectation that a participant is interested in Other Adult Materials. High frequencies of the People & Society, News, Law & Government, and Education topics reduces the model's expectation that a participant is interested in Other Adult Materials.

A.2 Impact of the Override List

Recall that Topics uses two approaches to map domains to topics, the first of which is an override list that is manually constructed by Google (see § 2). When the override list changes, this may alter the resulting distribution of topics produced by Topics, which might in turn impact the predictive power of models trained on topics data. To investigate this, we reran our Topics simulator while using three different override lists: no list at all (i.e., all domains are classified using the BERT model), the original version 1 override list, and the current version 2 list. For each scenario, we computed new topics distributions and trained new RF models. Figure 16 shows the results of these experiments: we do not observe noticeable performance differences across different versions of the override list.

To explain this result, we present Figure 17, which shows the empirical CDF of overlap between each participant's browsing history and each version of the override list. The solid lines present overlap when we consider the unique set of domains visited by each participant. The dashed lines present overlap when we weight each domain by visit frequency (on a per participant basis).

Demographic Category	Precision	Recall	F1	F1 (macro)
Less than $50,000	0.58	0.53	0.55	0.66
50,000–99,999	0.53	0.19	0.28	0.53
$100,000 or more	0.43	0.08	0.14	0.50
18-34 years	0.38	0.13	0.19	0.55
35-54 years	0.50	0.36	0.42	0.58
55 years or older	0.63	0.69	0.66	0.64
Male	0.62	0.61	0.61	0.63
Female	0.70	0.74	0.72	0.69
White	0.81	0.90	0.85	0.53
Black	0.17	0.17	0.17	0.55
Asian	0	0	0	0.50
Hispanic	0	0	0	0.49
Democrat	0.54	0.72	0.62	0.44
Republican	0.50	0.16	0.24	0.59
Independent	0	0	0	0.42
No College	0	0	0	0.47
Some College	0.64	0.73	0.68	0.55
Post-grad	0.25	0.02	0.04	0.43

Sensitive Interest	Precision	Recall	F1	F1 (macro)
Abortion	0.14	0.18	0.16	0.54
Alcohol	0.54	0.51	0.52	0.57
Folklore	0.30	0.22	0.26	0.56
Gambling	0.47	0.41	0.43	0.60
Global Religion	0.72	0.79	0.75	0.62
Health and Wellness	0.96	1.00	0.98	0.49
Marijuana	0.11	0.03	0.05	0.46
Medicine	0.72	0.83	0.77	0.62
Other Adult Materials	0.67	0.63	0.65	0.66
Political Organizations	0.81	0.99	0.89	0.45
Pornography	0.62	0.56	0.59	0.63
Tobacco	0.22	0.19	0.20	0.53
Weapons (Sales)	0.32	0.24	0.28	0.54

Table 5: Precision, recall, and F1 score for our demographic and sensitive interests prediction models. We present both binary and macro F1 scores.

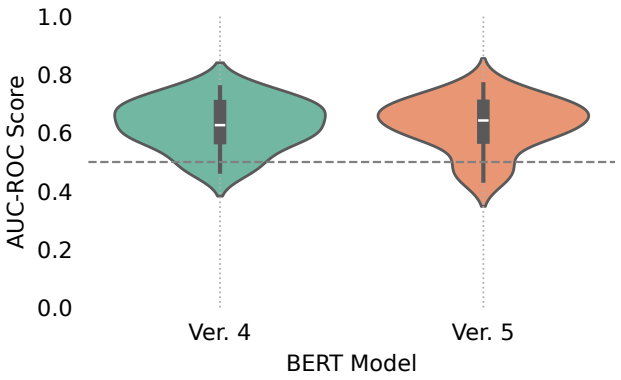

Figure 18: Distribution of AUC-ROC scores for all of our 31 models as we vary what BERT model version is used in Topics.

We observe that, in terms of unique domains, override list version 1 covers 18% of the median participant's browsing history, while override list version 2 covers 28%. However, in terms of weighted domains, override list version 1 covers 61% of the median participant's browsing history, while override list version 2 covers 69%. From this analysis we can conclude two things:

(1) Regardless of version, the override list covers the majority of participants' browsing history for the purpose of mapping topics. Recall that Topics chooses the most frequent topics per epoch, which are necessarily those topics mapped from frequently visited domains. Thus, for most participants, the composition of the override list is more important than the BERT model for determining the topics that they are assigned.

(2) Although override list version 2 includes an order of magnitude more domains than override list version 1, we observe diminishing returns in terms of the ratio of domains successfully mapped when moving from list 1 to list 2. This makes sense given that browsing behavior is highly skewed in favor of the biggest websites, and these websites were already included in override list version 1.

A.3 Impact of the BERT Model

The second approach Topics uses to map domains to topics is a BERT model. Like the override list, Google has updated the BERT classification model over time. Changes to the BERT model may also alter the distribution of topics produced by Topics. To investigate this, we reran our Topics simulator while using the two most recent (as of this writing) BERT models: versions 4 and 5. For each scenario, we computed new topics distributions and trained new RF models. Figure 16 shows the results of these experiments: we do not observe noticeable performance differences across versions 4 and 5 of the BERT model. This result is explained by our observation that the override list covers the majority of participants' browsing history (after accounting for visit frequency, see § A.2 and Figure 17).