Measuring Sensitivity of Cohorts Generated by the FLoC API

Andrés Muñoz Medina
Google Research

Michael Kleber
Google Chrome

Josh Karlin
Google Chrome

Marshall Vale
Google Chrome

Abstract

We present a discussion of the protections beyond k-anonymity that the Chrome implementation of the FLoC API will provide users. These protections mitigate the risk that a cohort number generated by this API in Chrome leaks sensitive information about the browsing behavior of a user.

Introduction

Today, many publishers are able to leverage interest-based advertising as a source of funding. This revenue stream allows them to offer content free of charge to their users. Contrary to contextual ads, interest-based ads leverage information about a user's interests to decide what ads to show them. Interest-based advertising enables an overall better ad experience for users because the user is presented with more relevant ads than traditional run of network ads; for advertisers, who can better reach their target audience; and for publishers, who are allowed to earn more money, on average, per interest-based ad than a non-relevant ad. In fact, multiple studies from academia and industry have consistently demonstrated that personalized advertising can account for 50-65% of a publisher's revenue.[1]

In order to accurately serve interest-based ads, ad tech companies use third-party cookies to generate user interest profiles. Thus, the planned deprecation of third-party cookies in Chrome puts interest-based ads, and the revenue publishers depend on, at risk. To ensure publishers continue to have options to fund their services, Chrome has proposed the FLoC API as a way to enable interest-based advertising in a private way. At a very high level, the FLoC API assigns users to a cohort in such a way that users in the same cohort have similar interests. An ad tech company can then use the API to advertise to an entire cohort.

---

[1] Beales, J. H. & Eisenach, J. A. (2014). An empirical analysis of the value of information sharing in the market for online content. Technical report, Navigant Economics.
Johnson, G., Shriver, S., & Du, S. (2020) Consumer privacy choice in online advertising: Who opts out and at what cost to industry? *Marketing Science*.
Ravichandran, D., & Korula, N. (Google 2019) "Effect of disabling third-party cookies on publisher revenue"

It has been shown that the FLoC API allows ad tech companies to enable interest-based advertising without generating fine-grained browsing profiles of users. The FLoC API achieves this by generating k-anonymous cohorts. That is, the API returns a cohort number shared by at least *k* users. This id can be used as an anonymous replacement of a third-party cookie, allowing ad tech companies to build *cohort interest profiles* without knowing the identity of a user.

While k-anonymity, especially for large values of *k*, protects users from reidentification, it is well known in the privacy community that this privacy notion can be vulnerable to so-called homogeneity attacks. In the context of the FLoC API, a homogeneity attack corresponds to a scenario where all users that share a cohort number also share a sensitive attribute. For instance, a cohort that consists only of users who visited a website about a rare medical condition. By revealing the cohort of a user, the FLoC API may inadvertently also reveal that a user has investigated that rare medical condition.

At a very high level, we want to make sure that no company, including Google, can correlate a particular cohort with any sensitive attribute.

The purpose of this paper is to discuss the privacy protections that are needed in order to prevent this type of privacy leakage and what Chrome is doing to prevent homogeneity attacks from happening in the initial FLoC API origin trial. As the implementation of the FLoC API is the responsibility of each browser or software that supports the API, the description of the protections here describe only the implementation by Chrome and not necessarily characteristics that are intrinsic to the API itself.

The sensitive cohort detection described below considers the risk that certain cohorts might imply an elevated likelihood of sensitive browsing behavior. There is a separate threat, not considered in this analysis, of an attacker attempting to guess browsing history based on the details of how cohorts are created. That risk should be mitigated by other measures designed to ensure that the map from browsing history to cohorts is sufficiently lossy, even when conditioned on other information a site might have about one of its visitors. Such measures warrant further investigation, but are out of scope for this document.

Sensitivity

Before describing the protections Chrome will put in place, we need to define what sensitive categories are. We will use the same sensitive interest categories defined by Google for its interest-based (personalized) advertising product[2]. This list of categories was chosen because Google already forbids showing ads related to them as well as targeting a user based on them. Examples of categories in this list are *adult* and *medical* websites as well as sites with *political* or *religious* content. We will use these categories to decide whether or not a web page is sensitive. While this list of categories certainly does not capture all the nuances of sensitive

[2]https://support.google.com/adspolicy/answer/143465?hl=en

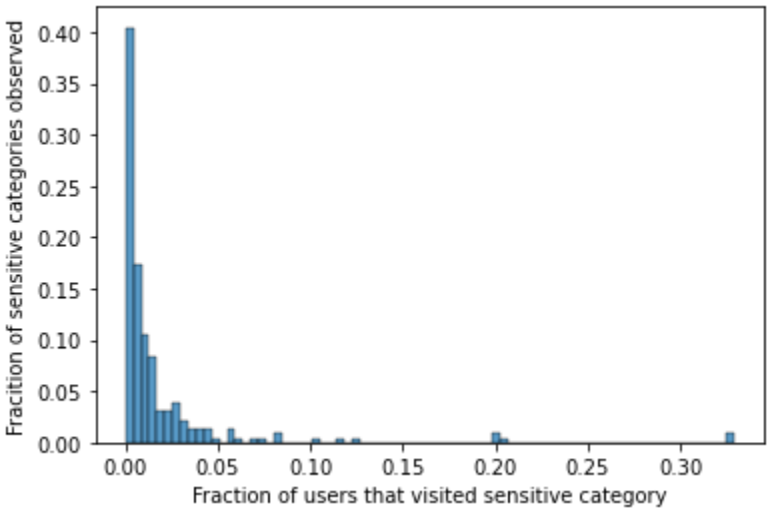
content (for instance, websites that are not sensitive but that a malicious actor might use, perhaps in combination with other data, as a proxy to infer sensitive attributes), we believe it provides us with a solid foundation that we can build upon. Moreover, the methodology presented here can be applied to any other ontology of sensitive categories as well.

Now that we have established what content is sensitive, we define how we decide whether a cohort leaks sensitive information or not. At a very high level, we want to ensure that no cohort consists of users that have visited web pages related to a particular sensitive category at a much higher rate than the general population. More formally, we ensure that a cohort assignment satisfies the strong privacy notion of *t-closeness*. A cohort assignment is said to satisfy *t-closeness* if it is k-anonymous and for every sensitive category, the distribution of users who visited a web page related to that category has distance at most *t* from the general distribution. Intuitively, *t-closeness* ensures that an adversary that observes a cohort number cannot infer much more about the sensitive browsing behavior of a user than they could before knowing their cohort. We now define the distance that will be used to implement *t-closeness.*

Definition 1. For each sensitive category *X* we define its population visit frequency as

PopulationFreq(X) = fraction of Chrome users that visited a website tagged with category *X*

We emphasize that a user is considered to have visited a sensitive category even if they visited only a single web page tagged with that category. The image below shows a histogram of the proportion of users who visited a given sensitive category. Notice that the majority of categories are visited only by a small fraction of users: more than 90% of the categories are visited by less than 5% of users.



Definition 2. For a cohort *C* and category *X* we define the cohort visit frequency as

CohortFreq(X, C) = fraction of users in *C* that visited a web page tagged with category *X*.

The goal of t-closeness is to ensure that the population and cohort visit frequencies are close for all cohorts and all categories.

Definition 3. We say category X^* is an anomalous category for a cohort *C* if it satisfies:

$$X^{*} = \arg\max_{X} \; CohortFreq(X, C) - PopulationFreq(X)$$

That is, the anomalous category is the category where the sensitivity in the cohort differs the most from the population sensitivity. We call the difference in sensitivity scores the sensitivity gap.

Definition 4. Sensitive cohorts. Given a threshold $t > 0$ we say a cohort is sensitive if the **sensitivity gap** of its anomalous category is larger than a threshold t.

For example, suppose t=0.1, and there is some sensitive category *X* that appears in the browsing history of 20% of the overall population. If that category appears in the browsing history of more than 30% of the people in some particular cohort, then the cohort is considered sensitive.

Notice that by definition of the anomalous category, every cohort that is not deemed sensitive satisfies *t-closeness.* When a cohort is sensitive, instead of returning the true cohort number, the FLoC API will return an empty string.

Implementation

We now go into detail of how we calculate the sensitive cohort in the initial FLoC origin trial. Currently this implementation depends on synced Chrome history data as the input to the batch generation of the sensitive cohort list that is eventually used in the FLoC API within the browser . This data source consists of Chrome users that satisfy the following conditions:

1. The user is logged into a Google account and opted to sync history data with Chrome
2. The user does not block third-party cookies
3. The user's Google Activity Controls have the following enabled:
    i. "Web & App Activity"
    ii. "Include Chrome history and activity from sites, apps, and devices that use Google services"
4. The user's Google Ad Settings have the following enabled:
    i. "Ad Personalization"
    ii. "Also use your activity & information from Google services to personalize ads on websites and apps that partner with Google to show ads."

The pipeline to calculate the sensitivity scores proceeds as follows:

- For each website in synced Chrome history data, we use an in-house classifier to obtain the sensitivity categories associated with the website.
- For each category *X* we calculate PopulationFreq(X) by counting the number of users who visited a website tagged with that category.
- For each category *X* and cohort *C* we calculate CohortFreq(X, C) similarly.
- For each cohort we decide whether the cohort is sensitive or not.
- We generate a blocklist of sensitive cohorts.
- This blocklist is updated every periodically to ensure we capture changes in browsing behavior.

While sensitive cohorts are identified using synced Chrome history data, the resulting list of sensitive cohorts is pushed to all instances of Chrome. In particular, the blocklist provides sensitivity protection to all users and not only those that activated Chrome Sync.

Use of synced Chrome history

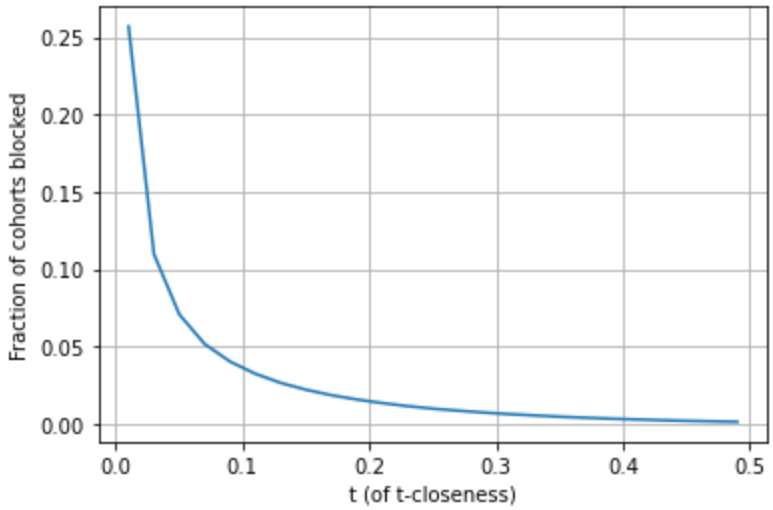
The use of synced Chrome history data is a preliminary solution for the initial origin trial; there is ongoing work to build alternative approaches that do not rely on this data source. Given that the cohort sensitivity can be calculated by counting the number of users in a cohort that visited a sensitive category, we need a system that allows us to query this information privately across all browsers. This will become feasible once the aggregate reporting infrastructure of the Privacy Sandbox is implemented. We also need a way to recognize which pages are in each sensitive category, which may be possible using on-device classifiers instead of the existing server-side url-crawling approach. But all of these implementation variants still perform the same t-closeness calculation described above.

Privacy considerations

It is natural to ask whether calculating the sensitivity of the cohorts itself raises privacy concerns for users, since it requires access to their browsing history in the initial origin trial. To protect access to the raw data, we run all calculations involving unaggregated user information in memory and we never write this information to disk. Moreover, notice that the final output of the pipeline is a blocklist of sensitive cohorts. In particular, we only learn whether or not a cohort is in a sensitive category, and not which category it is in.

Utility vs privacy

We now discuss the utility-privacy tradeoff of blocking sensitive cohorts. The plot below shows on the x-axis the level t used to deem a cohort sensitive. The y-axis corresponds to the fraction of cohorts that would be blocked using that threshold. While the actual utility drop of blocking certain cohorts is hard to measure, based on a previous white paper showing the benefits of the cohort signal for interest-based advertising, we believe it is reasonable to assume that utility decays as we block more cohorts. Thus, we would like to block the smallest number of cohorts while providing strong privacy protections. This curve can help guide browsers on how to choose a threshold for blocking sensitive cohorts. For instance with t=0.05 we would drop roughly 7% of the cohorts, whereas for t=0.1 we would drop only 4%. On the other hand, setting t=0.025 would drop more than 15% of the cohorts. This suggests that a better privacy-utility trade-off can be found in the interval [0.05, 0.1] than in the interval [0.025, 0.05]. For the first origin trial, Chrome has chosen 0.1 as the sensitivity threshold.



The empty cohort

Another information leakage risk is that of the empty string returned by the FLoC API. In theory, all cohort numbers blocked could share the same anomalous category, thus creating a correlation between the blocked cohort and that anomalous category. To ensure that this is not the case, we calculated the sensitivity profile of all users that would have been blocked (using a threshold t = 0.05) and compared it to the population sensitivity profile. The histogram below shows the distribution in the difference between the sensitivity scores of each category. We can see that only two categories deviate from the population average by more than 0.02 and that this is still below the sensitivity threshold t = 0.05. In practice, this population will be further mixed with other groups that are not assigned to any cohort (incognito browsing, opt-out, recently cleared their cookies, etc.), providing additional protection.