Privacy Limitations of Interest-based Advertising on The Web: A Post-mortem Empirical Analysis of Google's FLoC

Alex Berke, Dan Calacci of MIT Media Lab

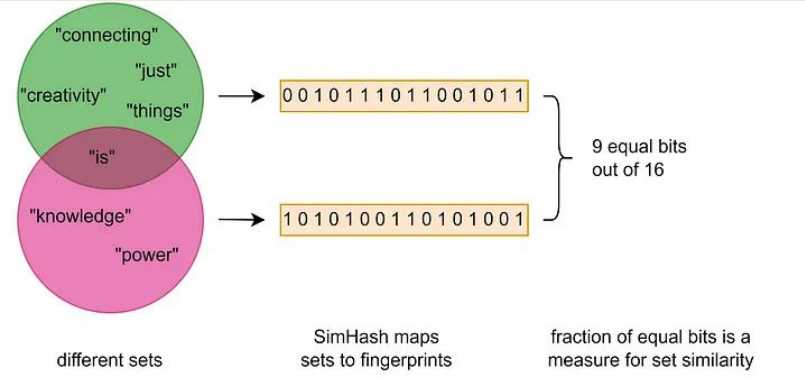
Discussion Presented by Emma Hartman

Introduction

Google's Privacy Problem

- Many browsers block third-party cookies by default
- Google owns both Google Chrome and Google Ads
  - Ads as a source of revenue
  - Behind in blocking third-party cookies
- Can Google have it all?
  - Block cookies and preserve privacy while still feeding relevant ads to Chrome users

Introducing Google FLoC

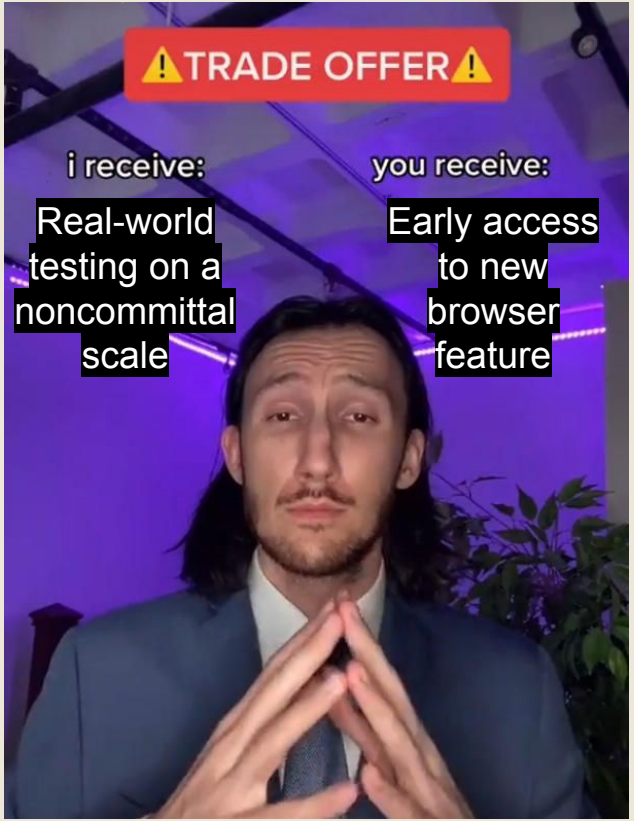


Source: https://medium.com/dynatrace-engineering/speeding-up-simhash-by-10x-using-a-bit-hack-e7b69e701624

- Federated Learning of Cohorts
- Input: public domains visited in the last week
- Output: SimHash bitvector
  - Locality-sensitive
- Hashes sorted into groups of size k using Google's PrefixLSH algorithm
  - Done centrally in an "anonymity server"
- Recomputed periodically

FLoC's Origin Trial

- Deployed to browsers and API provided to limited developers
- Chrome users with at least seven domains in browsing history
- Trial ran from Spring to Fall 2021
- Project cancelled in 2022



Source: bradeazy on TikTok

So What Happened to Google FLoC?

FLoC Had Some Haters

- Mozilla report suggested that FLoC might still enable user tracking
  - Despite only two cohorts, 6 users can be uniquely identified
- Other privacy advocates concerned that FLoC could reveal sensitive information about cohorts
- No empirical evaluation by Google or other parties on these claims

	User device 1	User device 2	User device 3	User device 4	User device 5	User device 6
Fingerprinting data	A	A	B	B	C	C
Period 1 cohort ID	1	1	1	2	2	2
Period 2 cohort ID	1	2	2	1	1	2
Period 3 cohort ID	1	2	1	1	2	2

Source: Privacy Limitations of Interest-based Advertising on The Web: A Post-mortem Empirical Analysis of Google's FLoC

Two Big Questions

1. Can individual users be identified on a larger scale?
    a. With and without fingerprinting
2. Will cohorts of specific sensitive topics be created?

Methodology and Evaluation

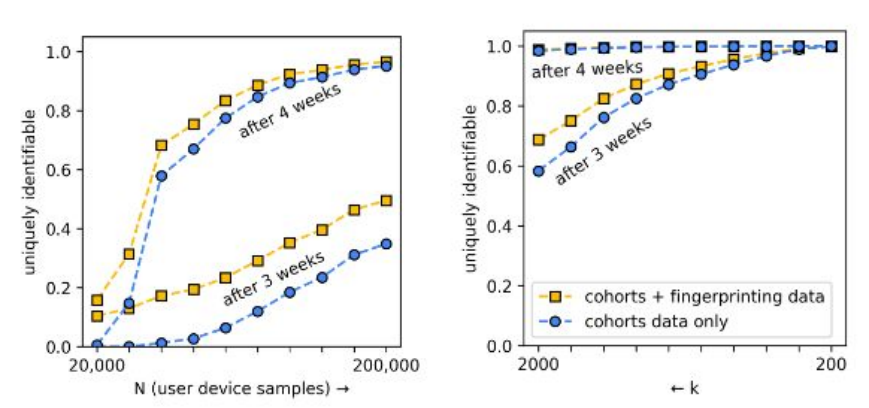
Recreating the Origin Trial

- Browsing data from comScore Web Behavior database
  - 50,000 households, 90,000 devices
  - All 52 weeks of 2017 counted
  - Self-reported demographics
- Searches sorted by week
  - Records with fewer than 7 unique domains dropped like the original
- Open-source SimHash verified by Google engineers used to implement PrefixLSH and compute cohort IDs
  - Assume FLoC cohorts are recomputed every 7 days
- Dataset expansion by splitting users into 13 4 week sequences
  - t-closeness and Pearson score verify similarity to the U.S. population

Machine ID	Session ID	Duration	Domain	Pages	Date	Time	...	Household Income	Race	Zip
169007206	19308896	33	site.biz	2	20170515	7:25:23		14	1	36832
169007206	27157206	5	example.com	1	20170515	8:36:55		14	1	36832
170422065	67238569	46	google.com	3	20170515	23:27:22		16	1	80233

Source: Privacy Limitations of Interest-based Advertising on The Web: A Post-mortem Empirical Analysis of Google's FLoC

User Identification



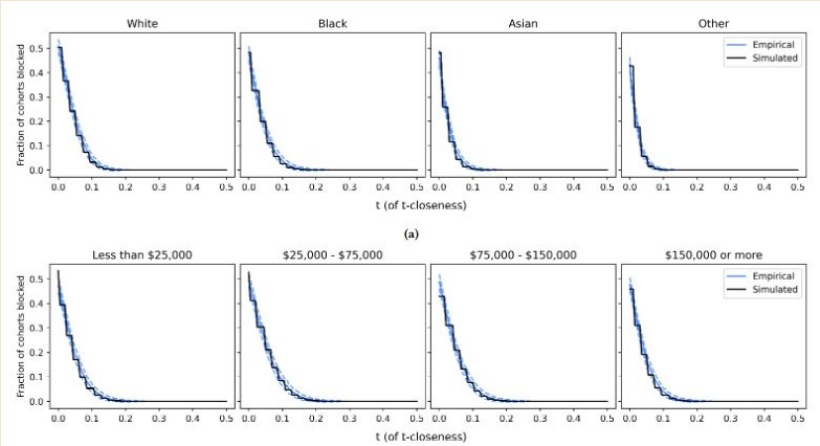
Source: Privacy Limitations of Interest-based Advertising on The Web: A Post-mortem Empirical Analysis of Google's FLoC

- Over 50% of samples uniquely identifiable after week 3
  - Over 95% by week 4
- Weak fingerprinting increases risks of unique user identification
- Conservative underestimate of the risk
  - Smaller sample size so less cohorts to identify unique traits from

Demographic Identification Setup

- Race and income considered
- t-closeness formula Google uses to filter demographics used to determine cohorts that group together those of sensitive demographics
- "Panels" of proportionally selected users to minimize bias for the individual sample
  - Smaller cohort size to account for smaller datasets
- Compare to randomized SimHashes and randomly assigned race and income

Demographic Identification Results



Source: Privacy Limitations of Interest-based Advertising on The Web: A Post-mortem Empirical Analysis of Google's FLoC

- Browsing behaviors do differ by race and income on a significant level
- Despite differing browsing behavior, cohort violations of t-closeness are equal to random chance
- Three presented reasons
  - Browsing patterns weren't sufficiently different between these groups
  - FLoC finds stronger patterns elsewhere
  - FLoC is not good at what it does

Summary of Conclusions

- Individuals can be uniquely identified after multiple rounds of cohort placement
- Despite unique internet searches, race and class demographics are not grouped together to a significant degree
- The authors recommend contextual ads via manual purchasing per website and that FLoC-like projects of the future should analyze effects on sensitive demographics

Your Thoughts

What You Liked

- First to give a proper empirical analysis of these risks in FLoC
- Interesting result!
- Large dataset used with attempts to account for size
  - Pearson correlation and t-closeness analysis shows that the population is similar to the U.S.
- Made the limitations of their dataset clear and justified scaling methods
  - Underestimate given!

What You Wanted Improved

- Very limited demographic characteristics, look into other demographics
- The dataset
  - Demographics were self reported
  - 50,000 users
- No qualitative comparison to Google Topics
- Recommendations left something to be desired

Questions and Other Observations

- Are FLoC's privacy issues fixable or is the concept inherently flawed?
  - Was FLoC specifically salvageable?
  - Are cohort-focused privacy projects generally viable without violating privacy?
- How does FLoC compare to Topics?
- Google's reasoning for ending FLoC was not very insightful...so let's speculate.
  - Why do you think they cancelled the project?
  - Why were they vague about their reasoning for ending the project?
- What is the point of a paper like this?

Your Ratings