Replication: Why We Still Can't Browse in Peace: On the Uniqueness and Reidentifiability of Web Browsing Histories

Sarah Bird, Mozilla
Ilana Segall, Mozilla
Martin Lopatka, Mozilla

Original Paper

Browsing history - the set of domains you have visited - is highly unique and could be used as a tracking vector.

Why replicate:

- The web and browsing has evolved: user generated content and core platforms
- Tracking ecosystem has grown and consolidated
- We can collect more detailed data to answer questions about reidentifiability raised by original paper

Why Johnny Can't Browse in Peace: On the Uniqueness of Web Browsing History Patterns

Olejnik, Castelluccia, and Janc

5th Workshop on Hot Topics in Privacy Enhancing Technologies (Hot-PETS 2012)

Background & Definitions

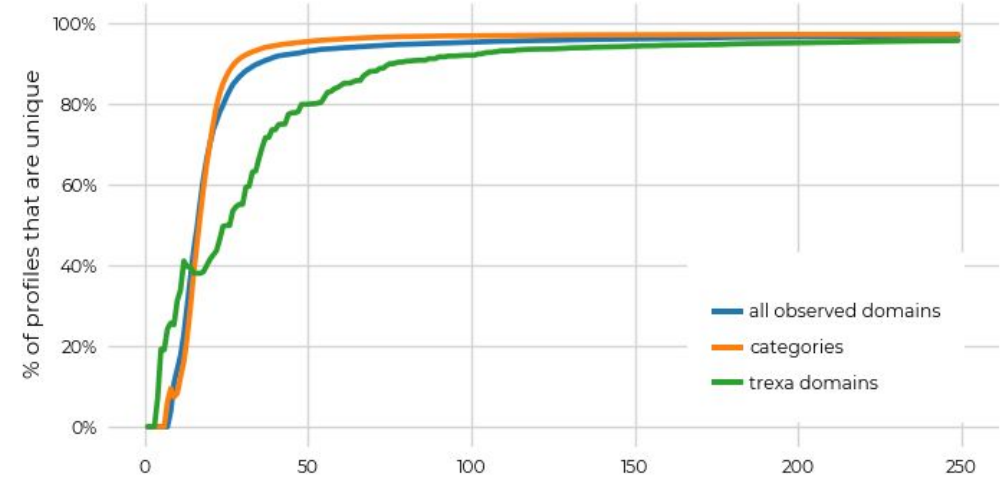
52,000 Firefox opt-in users

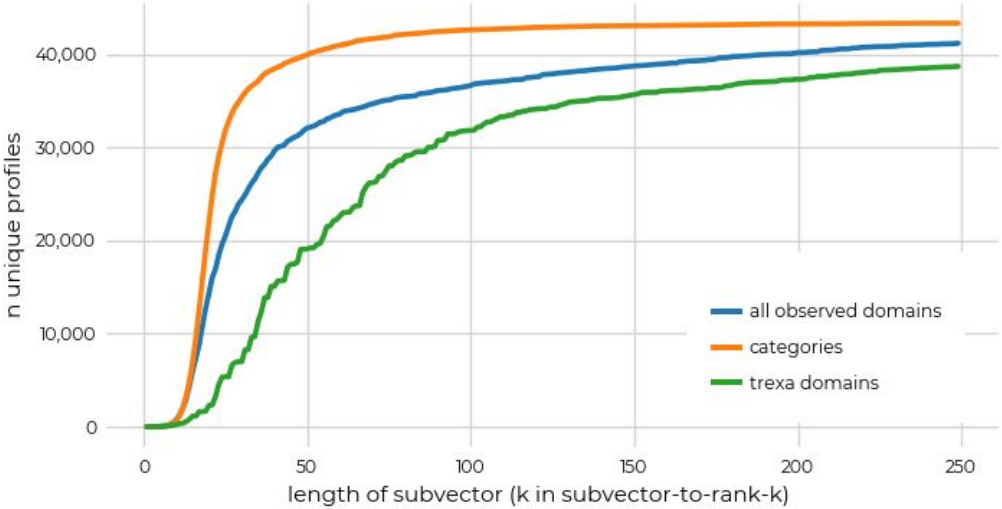
2 weeks of data collection

35 million site visits

660,000 distinct domains

- Types of profile
  - All observed domains
  - Predefined list of domains - Trexa (Tranco + Alexa)
  - Categories
- Profile - a list of x that a user visited
- Profile size - how many x did a user have in total?
- Length of subvector - how many x are we considering?





Replication

We replicate the core findings of Olejnik et al.

A large proportion of profiles are unique.

This holds even for small profiles e.g. 50 domains.

Extension

We move beyond profile stability to measure a reidentification rate.

Jaccard distance - degree of overlap between two sets.

(a) For each wk1 profile compute the Jaccard distance to all wk2 profiles

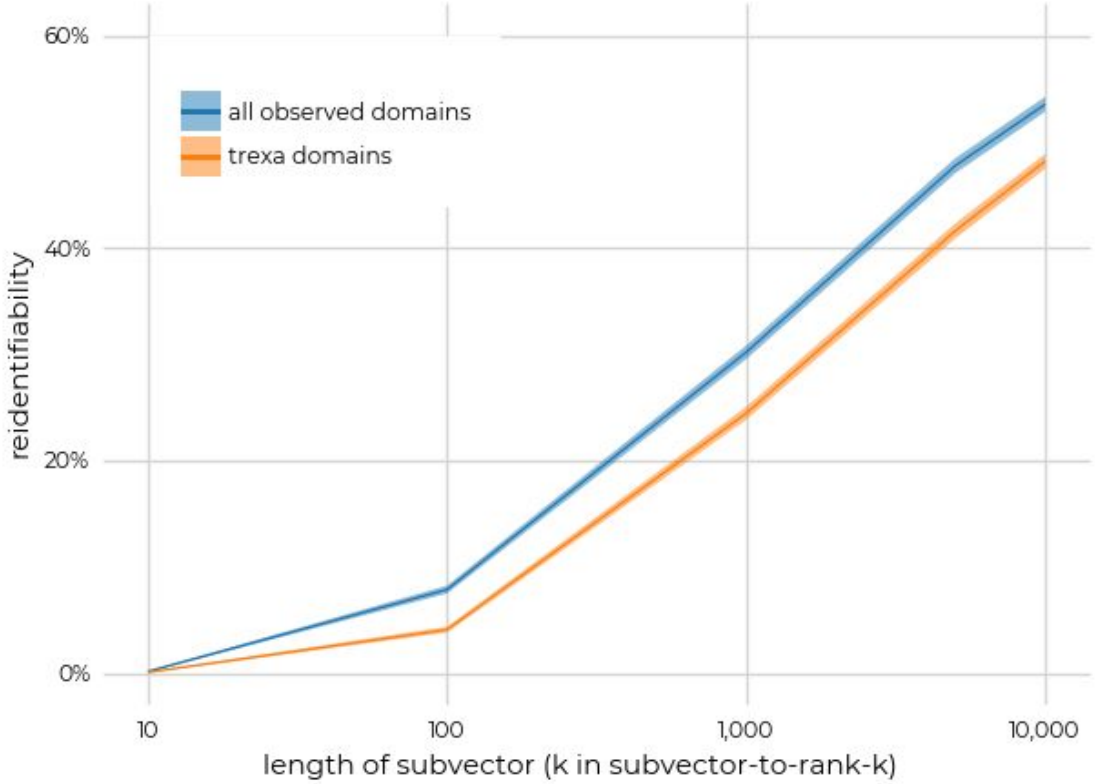
(b) pick the profile with the lowest Jaccard distance

(c) If wk1 and wk2 users are the same, it is a match.

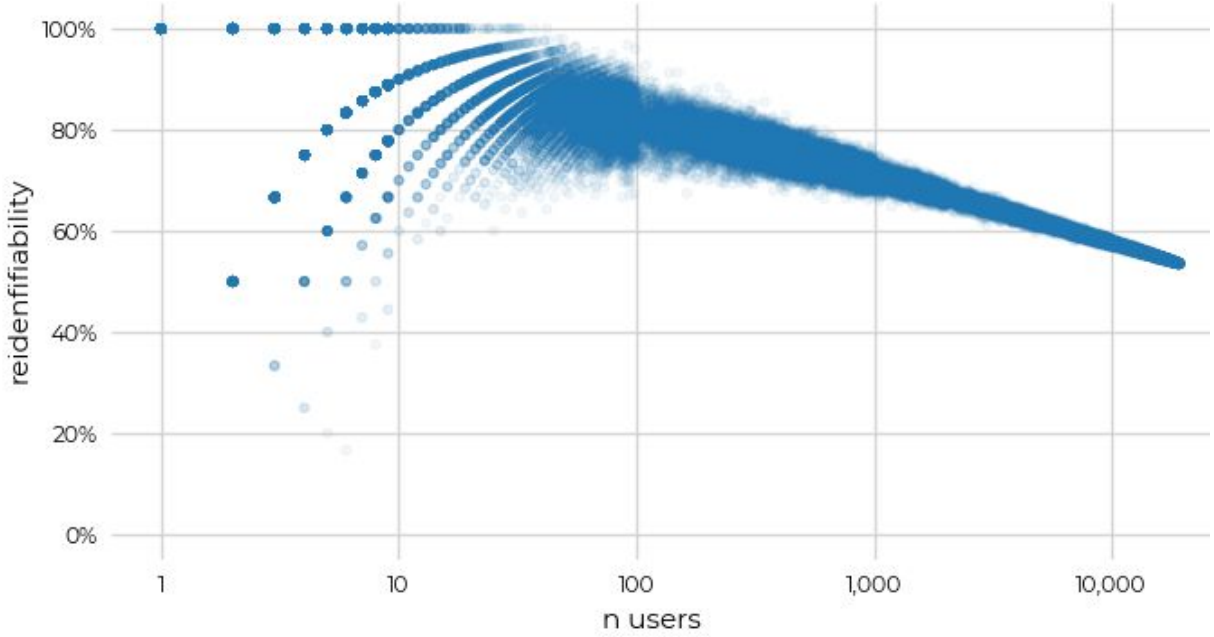
Reidentifiability metric: % of users correctly matched

Baseline reidentifiability

n = 19,263 users with profile size > 50



Scalability

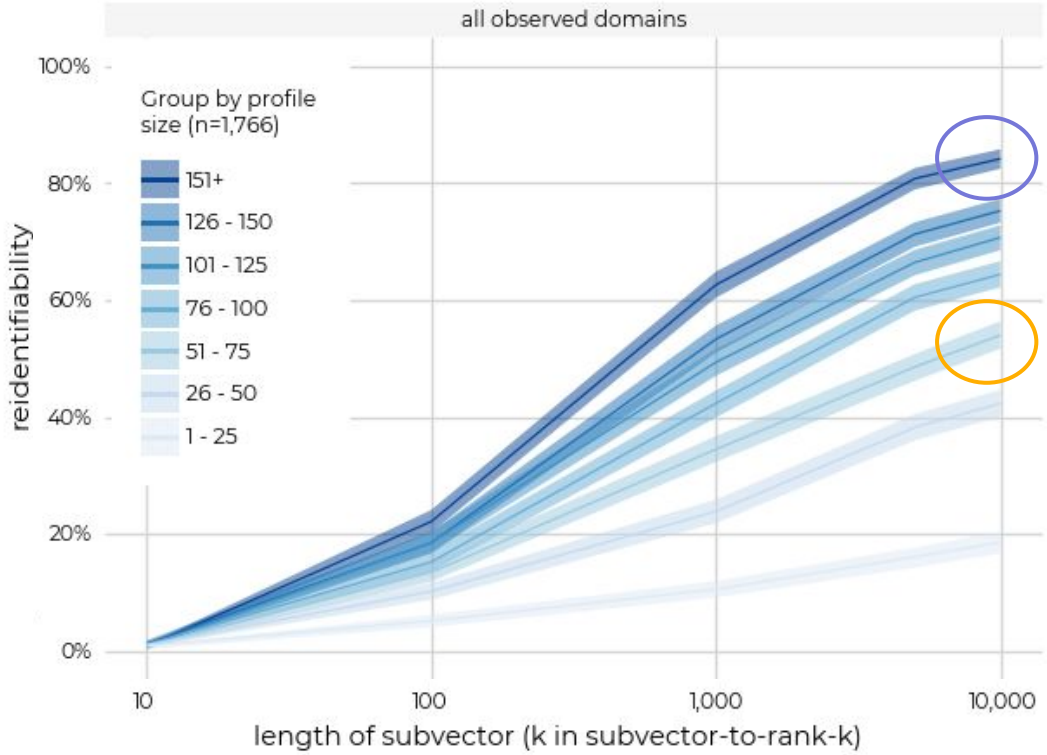


A 10x reduction in the number of users increases reidentification rate by 10%.

Monte Carlo simulation on users with profile >50, sampling between 1 and 19,263 users, 55,000 times.

Profile Size

~80% for profile size >150

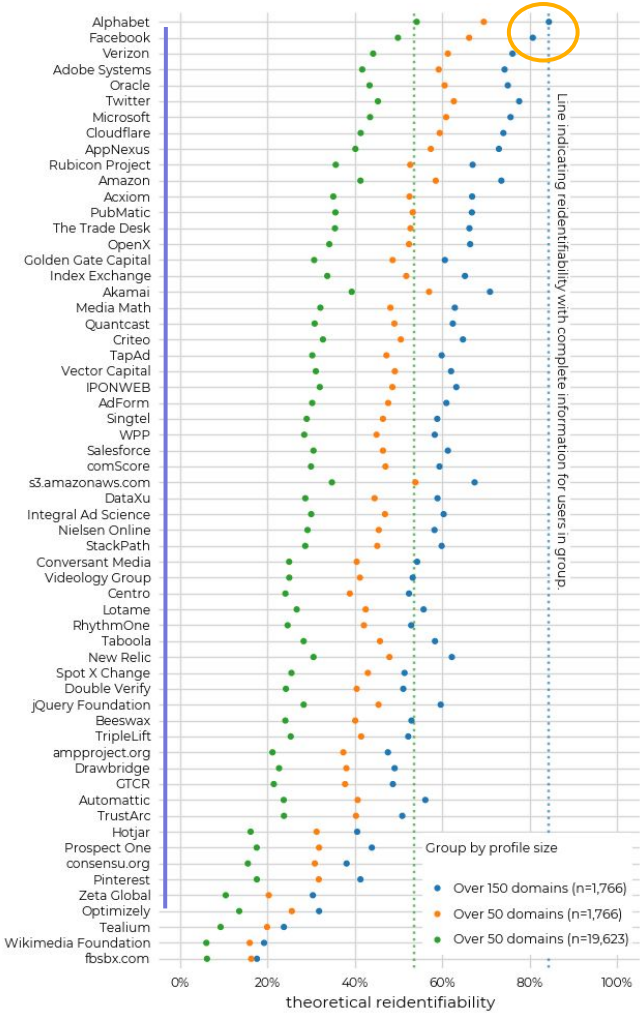


~50% for profile size ~50

Compute reidentifiability rates for equally sized groups of users (n=1,766) with different profile sizes.

Reidentifiability does not change dramatically between all domains and Trexa list.

Third-parties



Alphabet and Facebook have close to maximum theoretical reidentifiability rates.

A large number of third parties have sufficient presence for meaningful reidentification rates

Use complete request-response data to identify actual exposure to third parties (grouped by entity e.g. Alphabet parent company of Google and others).

Discussion!

Sarah Bird, Mozilla*
Ilana Segall, Mozilla
Martin Lopatka, Mozilla

* Corresponding author - sbird@mozilla.com