The Uniqueness and Reidentifiability of Web Browsing Histories

Sarah Bird and Ilana Segall

Abstract:

We examine the threat to individuals' privacy based on the feasibility of reidentifying users through distinctive profiles of their browsing history visible to websites and third parties. This work replicates and extends the 2012 paper Why Johnny Can't Browse in Peace: On the Uniqueness of Web Browsing History Patterns. The original work demonstrated that browsing profiles are highly distinctive and stable. We reproduce those results and extend the original work to detail the privacy risk posed by the aggregation of browsing histories. Our dataset consists of two weeks of browsing data from ~52,000 Firefox users. Our work replicates the original paper's core findings by identifying 48,919 distinct browsing profiles, of which 99% are unique. High uniqueness holds even when histories are truncated to just 100 top sites. We then find that for users who visited 50 or more distinct domains in the two-week data collection period, ~50% can be reidentified using the top 10k sites. Reidentifiability rose to over 80% for users that browsed 150 or more distinct domains. Finally, we observe numerous third parties pervasive enough to gather web histories sufficient to leverage browsing history as an identifier.

Bio:

Sarah is a Staff Software Engineer on the Privacy and Security Products team at Mozilla. Prior to that she was a Research Engineer on the Firefox Machine Learning team. Her research focused on developing statistical techniques for detecting tracking technologies. Previously, Sarah was a software engineer engaged in building websites, tools for data science, and open data standards. Sarah holds masters' degrees in Technology and Policy from MIT and Mechanical Engineering from the University of Cambridge. https://www.linkedin.com/in/birdsarah