9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

Browsing behavior exposes identities on the Web

Keywords: browsing behavior | world wide web | online privacy | human behavior | tracking

Extended Abstract

How easy is it to uniquely identify a person based on their web browsing behavior? Here we show that when people navigate the Web, their online traces produce fingerprints that identify them. By merely knowing their most visited web domains, five data points are enough to identify 95% of the individuals. These digital fingerprints are stable and render high re-identifiability. We demonstrate that we can re-identify 89% of the individuals in separate time slices of data. Such a privacy threat persists even with limited information about individuals' browsing behavior, reinforcing existing concerns around online privacy.

Context

In the era of ubiquitous technology, our online habits have become a goldmine for companies seeking to extract value from our data [1]. By looking at how we browse the Web, companies learn about ourselves, enabling them to build highly targeted advertising campaigns and create user profiles that can be sold to third-party advertisers [2, 3]. Their business model essentially centers on tracking and predicting individuals' behavior [1].

Recently, researchers have shown that our online behavior is indeed highly predictable, resulting from our tendency to follow web routines—our online habits make use predictable [4]. This predictability is not limited to online activity but extends to other domains, including mobility and shopping, which researchers have demonstrated that individuals can be uniquely identified based on their data [5, 6]. In this regard, previous works have examined the uniqueness of online traces from a technical perspective [7–9]. In contrast, in this work, we characterize the fundamental features in our behavior that expose our identities on the Web.

Results

To understand how online browsing behavior produces fingerprints on the Web, we analyze individuals' most visited domains, aiming to explore their ability to distinguish individuals apart. We demonstrate that we can uniquely identify a person by using a list containing their favorite domains. First, we describe each individual by a n-tuple consisting of their n most visited domains, and then we count the number of individuals having the same n-tuple (see Fig. 1). Our results reveal that most individuals have unique fingerprints when $n \geq 5$, implying that an individual's five most visited domains uniquely identify them. This short-length fingerprint holds true regardless of an individual's demographic profile, as we find no significant statistical differences in fingerprint lengths when accounting for individuals' gender or age.

These fingerprints could be even shorter if it were not for popular domains—individuals often have the same domains in their most visited list. For instance, 80% of the users have the same favorite domain as other users. These popular domains enlarge the fingerprints, whereas unpopular domains make them distinctive. For example, when we use the least visited domains to form the n-tuples, we find that most individuals have unique fingerprints when $n \geq 2$, implying that the two least visited domains of a user uniquely identify them. However, we expect considerable fluctuations in the list of the least visited domains over time, which raises the question of re-identifiability power based on these fingerprints.

To understand fingerprints' ability to identify individuals on the Web, we examine different time slices of data to determine the re-identification rate based on the fingerprints. We find that we can re-identify 89% of the individuals in separate time slices of data. With this work, we characterize the fundamental features of our behavior that reveal our identities on the Web, which provides us insights to develop better strategies to safeguard people's privacy and security online.

Data

Our anonymized dataset consists of one month's (October 2018) web tracking data of 2,148 German users, which forms a representative sample of German online population (under the age of 65). For each user, the data contains the URL of the webpage the user visited, the domain of the webpage, time of visit, and active seconds spent by the user on the page. In total, these 2,148 users made 9,151,243 URL visits, spanning 49,918 unique domains.

References

[1] Zuboff, S. Surveillance Capitalism and the Challenge of Collective Action. *New Labor Forum* **28**, 10–29 (2019).

[2] White, C. L. & Boatwright, B. Social media ethics in the data economy: Issues of social responsibility for using Facebook for public relations. *Public Relations Review* **46**, 101980 (2020).

[3] Cotter, K., Medeiros, M., Pak, C. & Thorson, K. "Reach the right people": The politics of "interests" in Facebook's classification system for ad targeting. *Big Data & Society* **8**, 205395172199604 (2021).

[4] Kulshrestha, J., Oliveira, M., Karaçalık, O., Bonnay, D. & Wagner, C. Web Routineness and Limits of Predictability: Investigating Demographic and Behavioral Differences Using Web Tracking Data. *Proceedings of the International AAAI Conference on Web and Social Media* **15**, 327–338 (2021).

[5] de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M. & Blondel, V. D. Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports* **3**, 1376 (2013).

[6] de Montjoye, Y.-A., Radaelli, L., Singh, V. K. & Pentland, A. S. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science* **347**, 536–539 (2015).

[7] Olejnik, L., Castelluccia, C. & Janc, A. Why Johnny Can't Browse in Peace: On the Uniqueness of Web Browsing History Patterns. In *Proceedings of the 5th Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs 2012)*, 17 (Vigo, Spain, 2012).

[8] Gómez-Boix, A., Laperdrix, P. & Baudry, B. Hiding in the Crowd: an Analysis of the Effectiveness of Browser Fingerprinting at Large Scale. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, 309–318 (ACM Press, Lyon, France, 2018).

[9] Bird, S., Segall, I. & Lopatka, M. Replication: Why We Still Can't Browse in Peace: On the Uniqueness and Reidentifiability of Web Browsing Histories. In *Proceedings of the Sixteenth Symposium on Usable Privacy and Security*, 16 (2020).

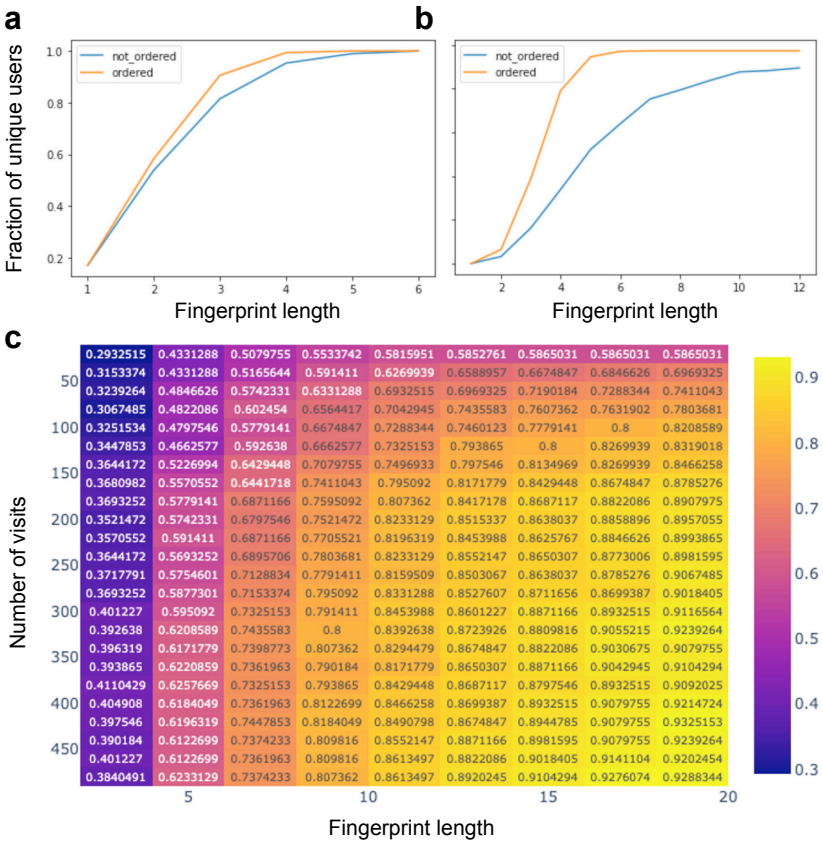

Figure 1: **People have preferences on the Web which distinguish them apart.** **(a)** We describe each individual by a n-tuple consisting of their n most visited domains, and then we count the number of individuals having the same n-tuple, revealing that most individuals have unique fingerprints when $n \geq 5$. **(b)** This uniqueness occurs even when we limit the number of domains to the Top 30 in the data set. **(c)** This unicity translates into re-identification, depending on the amaount of data we have for each user.