HAL
archives-ouvertes.fr

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

Lukasz Olejnik, Claude Castelluccia, Artur Janc

▶ **To cite this version:**

Lukasz Olejnik, Claude Castelluccia, Artur Janc. Why Johnny Can't Browse in Peace: On the Uniqueness of Web Browsing History Patterns. 5th Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs 2012), Jul 2012, Vigo, Spain. 2012. <hal-00747841>

HAL Id: hal-00747841
https://hal.inria.fr/hal-00747841

Submitted on 2 Nov 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

Łukasz Olejnik[1], Claude Castelluccia[2], Artur Janc[3]

[1] INRIA, Grenoble, France, lukasz.olejnik@inria.fr

[2] INRIA, Grenoble, France, claude.castelluccia@inria.fr

[3] Google, Inc., Mountain View, USA, aaj@google.com

Abstract

We present the results of the first large-scale study of the uniqueness of Web browsing histories, gathered from a total of $368,284$ Internet users who visited a history detection demonstration website. Our results show that for a majority of users (69%), the browsing history is unique and that users for whom we could detect at least 4 visited websites were uniquely identified by their histories in 97% of cases. We observe a significant rate of stability in browser history fingerprints: for repeat visitors, 38% of fingerprints are identical over time, and differing ones were correlated with original history contents, indicating static browsing preferences (for history subvectors of size 50). We report a striking result that it is enough to test for a small number of pages in order to both enumerate users' interests and perform an efficient and unique behavioral fingerprint; we show that testing 50 web pages is enough to fingerprint 42% of users in our database, increasing to 70% with 500 web pages. Finally, we show that indirect history data, such as information about *categories* of visited websites can also be effective in fingerprinting users, and that similar fingerprinting can be performed by common script providers such as Google or Facebook.

1 Introduction

Motivations: A body of prior work has studied the leakage of private information when users browse the Web. In addition to data obtained by direct observation of Web traffic, known vectors for privacy loss range from explicit inclusion of third-party scripts [11] to long-standing browser mechanisms which allow Web authors to query the contents of a user's DNS or browser caches [6, 25], history store [9] or configuration information [5].

In this paper we analyze the consequences of the existing ability of Web authors to determine which websites a user has visited. Specifically, we investigate whether a user's browsing history, i.e. the list of visited websites, constitutes a fingerprint which can be used to uniquely identify and/or track the user. As users' Web browsing preferences are directly related to the content in which they are interested, it is suspected that such preferences may be individual in nature, akin to traditional biometric mechanisms such as friction ridges of a finger, or retinal patterns.

Since, as we discuss in Section 2.1, there exist several largely reliable methods to determine if a user has visited a particular website, we expect that such knowledge can be easily gathered by invasive webmasters [10]. Therefore, a careful study of the consequences of such history detection from a privacy perspective is necessary.

Contributions: Our investigation is based on the analysis of a large dataset of $382,269$ users' browsing histories obtained in a real-world scenario [8]—each browsing history is a subset of websites visited by a given user. We also convert these histories into interest profiles by labeling each history element with a

category obtained by querying a website categorization service [17]. The interest profile of a user is then defined as the categories of the sites he/she visited.

We then analyze these two datasets, focusing on the following questions:

- How are web histories and interest profiles distributed? Are they unique or similar? How large is the set of visited websites which must be queried to accurately distinguish between users if at all possible?
- Are web histories and interest profiles stable? In other words, do they constitute good behavioral fingerprints and can tracking websites rely on such data?
- Can web histories and interest profiles collected by service providers, such as Facebook or Google, be used to fingerprint users in the absence of any other history detection or tracking mechanisms?

Our findings indicate that, under the sample we studied, the vast number of users' Web histories are distinct and in fact unique; and that this is still the case when analyzing the general categories of visited website sets, rather than individual websites themselves. Strikingly, it is also possible to attribute a distinct history fingerprint to a user by testing just a small number of pre-defined popular pages and such fingerprints, to some extent, are stable over time; in certain cases such detected browsing history sets are recreated after a one-time clearing of the browsing history.

The potential risks and privacy implications result from a combination of sensitivity and persistence of this type of interest-based behavior data. In general, it is not simple to change one's browsing habits—if Web browsing patterns are unique for a given user, history analysis can potentially identify the same user across multiple Web browsers, devices, or if the user permanently changes her physical location.

Paper organization: This paper is organized as follows. First, we provide an overview of existing methods which allow the detection of users' browsing histories, and prior work on Web-based fingerprinting techniques. We then outline our methodology for gathering and processing browsing history data. In Section 4.1 we analyze the distinctiveness of browsing profiles. In Section 4.2 we perform an analogous analysis using only high-level website categories. In Section 4.3 we review the stability of users' browsing histories. We conclude by outlining some countermeasures to history detection and analyzing the history tracking potential of third-party script providers such as Google and Facebook (Section 5).

Ethical considerations: In this study, we utilize data gathered by a proof-of-concept Web application which was created to inform users about the risks of history detection and describe mitigations [8]. Users visiting the site automatically executed the default history test and the detected contents of their browsing histories were sent to a server so that it could display all gathered information back to the user. As such, the dataset contains real history information including records of user visits to potentially sensitive websites which might, in certain cases, be harmful to users if revealed to other individuals or organizations (e.g. employers).

Recognizing the significant problem which arises from gathering such information, we have taken the following precautions when analyzing and storing data:

- Apart from showing each user the contents of their browsing history as detected by our system, all data was analyzed and displayed in aggregate, without uncovering any user-identifiable information.
- The system does not allow any users to view any past history detection results (including their own) and does not use any mechanisms for tracking users (i.e. cookies)
- All log data was deleted from Internet-facing hosts and was used solely to prepare aggregate information presented in this work.

We also believe that an important consideration is the fact that such usage data is obtainable by any website visited by the user, and the detection techniques are widely known. In fact, we are aware of several

toolkits which gather user history data and send them covertly to their origin websites [10]; the desire to understand the implications of such privacy leaks is a major motivation for this work.

2 Background

2.1 History detection mechanisms

Web authors have in their arsenal a variety of techniques which enable them to query the contents of a visitor's browsing history. One of the most well-known approaches, and the one used to gather data for this work, is the querying of URLs in the browser's history using the CSS :visited mechanism [9]. While modern browsers introduced fixes [2] for high-speed history detection via this technique, certain interactive attacks are still possible [19]. In addition, about 25% desktop user agents and an even higher proportion of mobile browsers are still susceptible to this technique [18].

An alternative approach is the timing analysis allowing detection of items in a Web browser's cache, introduced by Felten et al. [6] and recently perfected by Zalewski [25]. Yet another history detection vector is the timing of DNS queries [12, 20]. Even in the absence of such client-side timing attacks, in many cases it is possible to reveal if a user is logged into a particular website by analyzing error messages or timing server responses ; however, this technique is likely more difficult to generalize in a real attack.

Such history detection attacks have been successfully demonstrated in various scenarios. Wondracek et al. showed the potential to deanonymize social network users [21]. Jang et al. demonstrated that such techniques are indeed in use in the wild as well as study different leakage channels [10] which only makes this threat more significant. The work done in [9] revealed the susceptibility of the majority of Internet users to high-speed history detection and showed that it can uncover large amounts of data about users. Timing analysis techniques have been successfully demonstrated [25] and were practically used to discover the contents of the users shopping carts [3].

In addition to potentially being leaked to third-party Web authors, such Web browsing data is revealed to legitimate service providers such as DNS server operators, third-party script providers [11] or ad networks, even if it is not explicitly gathered. Thus, we expect information about websites visited by Web users as part of day-to-day browsing is quite easily obtainable by a variety of parties.

2.2 Web fingerprinting

There are few results on behavioral profiling and fingerprinting analysis based on large samples of data or results of real-world surveys. One recent and prominent is an excellent study of browser fingerprints in [5], where the fingerprinting is based on plugins, fonts and other browser configuration.

Another recent and important example is [24], where the authors study a large data sample from users of Hotmail and Bing and focus on the potential of tracking relating only to the host information and other such as browser cookies and User-Agent string. Fingerprinting potential based on the detection of browsers' configuration using JavaScript is also analyzed by Mowery et al. [15]; similar techniques have been employed by Eckersley in his experiment [5].

Different aspects of timing, as well as DNS cache timing attack are explored by Jackson et al. [7]. If the attacker has access to the routing nodes, he can use the network flow to fingerprint the users, as shown in [23]. Behavioral biometry, where fingerprints are based on the behavioral aspects and traits such as typing dynamics or voice analysis are described in [14, 13, 22]

3 Methodology

Data analyzed in this paper was gathered in the What The Internet Knows About You project [8], aimed at educating users and browser vendors about the consequences of Web browser history detection. For the overall discussion of the system refer to [9] where the authors discuss the susceptibility of Web users to such techniques, their performance, mitigations, as well as give a detailed review of history detection mechanisms.

3.1 Experimental Setup

The experimental system utilized the CSS :visited history detection vector [4] to obtain bitwise answers about the existence of a particular URL in a Web browser's history store for a set of known URLs.

The system leveraged a two-tiered architecture to first detect the "primary links", i.e. top-level links such as `www.google.com`, and use that knowledge to query for secondary resources associated with the detected primary link, such as subpages within each site.

The history detection demonstration site gained popularity and between January 2009 and May 2011 we gathered $441,627$ profiles of unique users who executed a total of $988,128$ tests. We expect that our data sample comes from largely self-selected audience skewed towards more technical users, who likely browse the Web more often than casual Internet users.

In this paper we refer to $382,269$ users who executed the default "popular sites" test of over 6,000 most common Internet destinations.

The "popular sites" list was created out of 500 most popular links from Alexa [1], $4,000$ from the Quantcast popular websites list [16], lists of common government and military websites, and several custom-chosen URLs selected for their demonstration and education potential. In this work we analyze data about visited "primary links" only, without analyzing any detected subresources within a website.

This approach did not make it possible to obtain user's entire browsing history, and that was not the aim of the study. However, this was not necessary, as it is sufficient to show that the actual subsets are unique: a considerable unique number of profiles within a large dataset most likely indicates the overall uniqueness of the whole history superset of these Web users (if the subsets of some supersets are unique, those supersets must consequently be unique).

Other history detection techniques we describe will likely have different semantics and capabilities, but will, in the majority of cases, be able to obtain similar answers about whether a user had visited a particular website. Thus, we can analyze the gathered data without loss of generality.

3.2 Data Collection

For each user, the set of all visited links found by the detection algorithm was sent to the server so that it could display results to the user—in the majority of cases, the detection phase took less than 2 seconds for the set of 6,000 popular sites.

In addition to history test results, the system gathered each user's browser version information (User-Agent header), IP address as visible to our webserver and the test date. The system did not make use of any side-information such as cookies, Flash cookies or other persistent storage techniques to track users and did not provide any capability to review any history results except for the most recent test run.

An important aspect of the system was educating users about the privacy risks associated with history detection; along with the results page listing detected URLs we provided users with general information about the problem, references to related research, and mitigations against history detection. Thus, it was assumed that a considerable number of users will clear their histories after visiting the site, so in our data

analysis we refer to history data received during the first test execution for a given user (except for the stability analysis of repeat visitors which analyzes data from subsequent test runs).

3.3 Terminology

The data we are summarizing in this work may be perceived as a collection of Web history profiles H_s (for sites). If a profile p_i, which is a collection of visited sites, is present in H_s, then $Count(p_i)$ is the number of occurrences of p_i in H_s. If $Count(p_i) = 1$, this profile is unique in a dataset (it is present only once, and thus relates to a unique user). When the number of detected links for a particular user is very small (1-3), it is likely that $Count(p_i)$ may be greater than zero as an obvious consequence of the pigeonhole principle. If $Count(p_i) > 1$ then this profile is not unique but we may treat is as distinct in the dataset.

Sites on the Internet may be attributed to different categories, for example `www.google.com` is a search engine, and `cnn.com` is a news portal. Therefore the sites in our dataset can be attributed to different categories and it is straightforward to define a category profiles collection H_c, where the Web pages are mapped to categories.

Moreover, H_s may be converted into a bit vector collection V_s in the following way: create a list of sites ordering them by popularity in the discussed dataset and map every profile $p_i \in H_s$ to a vector $v_i \in V_s$; a bit in this vector is set if a page on a given position is visited (present in p_i). Such conversion allows us to operate on bit slices of these vectors in the rest of our analysis.

4 Web Behavioral Fingerprints

To establish the potential of Web history traces as a fingerprinting mechanism we perform the analysis at three different levels of granularity. First, we analyze the raw data which includes all websites we detected in a user's browsing history. Second, we convert the raw history profile to a category profile where each website is assigned to one of several buckets (e.g. shopping, news, or social networking). Third, we convert such a category vector to a tracking vector as visible to scripts downloaded from the `facebook.com` and `google.com` domains (Section 5).

At each point we examine the stability of the resulting fingerprints.

4.1 Web History Profile Uniqueness

4.1.1 Methodology

We analyzed history profiles by computing distributions of the profiles per size of the profile (number of visited sites) to inspect profile diversity, both for all users and specifically for mobile browser users. The similarity metric we utilize is the Jaccard Index. For two sets A, B the Jaccard Index is computed as $\frac{|A \cap B|}{|A \cup B|}$. Two sets are equal if Jaccard Index is 1, and they are certainly similar (correlated) if it is larger than 0.7.

4.1.2 Results

Figure 1 displays the distribution of Web history profiles according to their size (pink line). It also shows the distribution of unique (red line) and non-unique profiles (green line). A profile is unique if it is associated with a single user. The unique distribution was prepared by counting all unique history sets - if a Web history profile appeared more than once ($Count(p_i) > 1$ for $p_i \in H_s$, which is especially the case of history sizes 1-5) it is counted as a single fingerprint (it is distinct but has non-zero siblings, thus non-unique).

A profile is non-unique if it is common to several users; a percentage of unique profiles is also presented. This percentage is computed by dividing, for each profile size, the number of unique profiles by the total number of profiles.

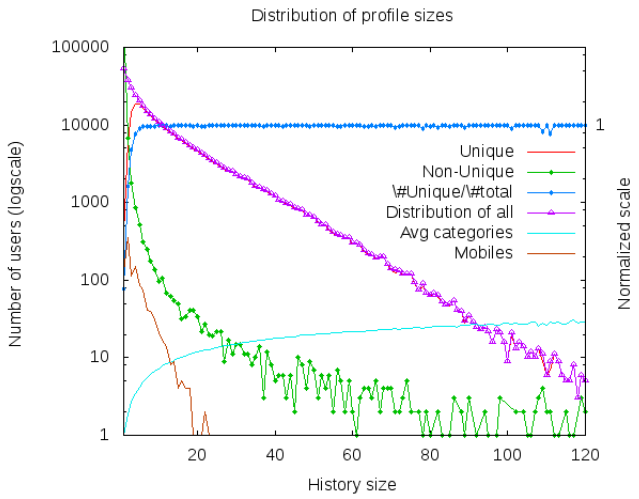


Figure 1: Distributions of: general, unique and non-unique Web history profiles per history size, normalized with respect to total number of histories in each history set. Average number of categories per history size is also shown.

It is interesting to note the clearly visible peak of detected sites - there are far more profiles with detected history size greater than 3, compared to smaller history sizes. The general distribution is also shown; it is easy to observe the prevailing unique profiles starting close to the point $X = 4$, indicating that detecting as few as four visited sites provides a useful fingerpriting signal.

In our dataset, the average number of visited sites per profile is 15, and the median is 10. However, analyzing just the history sizes larger than 4 ($223,197$ of such profiles) results in the average number of links 18 (median 13) with 98% profiles being unique.

For history sizes $1-4$ the ratio of unique profiles to total profiles ranges from 0.008 to 0.76 but in average, for all the profiles, 94% of users had unique browsing histories. This result suggests that browsing patterns are, in most circumstances, individual to a given user.

Figure 1 also displays the distribution of the Web history of mobile users (orange lines). The patterns of mobile Web usage is created by looking for a specific User-Agent HTTP header, detecting Web browsers on iPhone, iPad, Android and BlackBerry devices. The graph presents data from 1256 users. Even though this was not a large sample, with respect to overall number of profiles, different usage patterns are observed—specifically, the detected history sizes are smaller, which might suggest that the Web use on mobile devices is not as frequent or large as it is with non-mobile Web browsing.

To understand relative differences in observed history profiles, we analyzed the similarity between the fingerprints from different users. Within all of the history sizes, we measured the similarity using Jaccard index for all distinct (rather than unique) Web history profile pairs, i.e. if a profile was seen several times it is being treated as a single fingerprint. The averaged (by total number within history size set) result as a function of the history size is presented in Figure 2. The fingerprints were largely dissimilar, which confirms that enumerating actual browsing interests is feasible.

4.1.3 Selection

For a given website, what is its importance when studying history fingerprint uniqueness? If a given Web page is not very popular, or not visited very often, what is the "share" it provides to the overall uniqueness of a profile? Due to the sparseness of the set of visited sites for most users, we hypothesize that the most

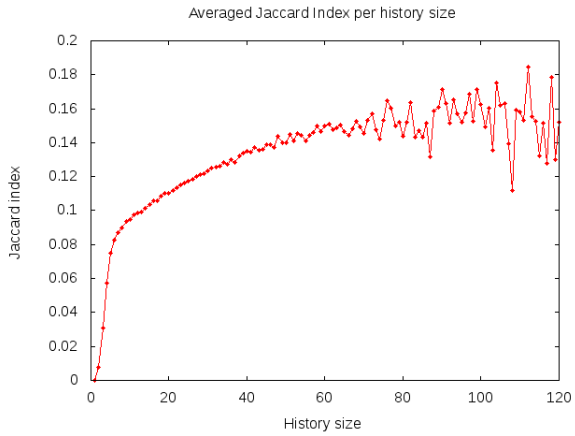


Figure 2: Jaccard Index as a function of history size. It is always smaller than 0.2 hence the profiles are dissimilar

commonly visited sites will be most useful in creating fingerprints. Of course, if a site is visited by only one person it contributes significantly to the uniqueness of this particular profile. However, it is of no value when establishing fingerprints of other users; it would likely negatively affect the performance of any real data-collecting implementation.

To analyze this, we sorted all of the websites with respect to their popularity metric in our data. After that, each Web history profile was converted to a vector representation. A bit in a vector was set if an associated page has been visited. Therefore, the most popular pages were the left-most bits. We studied slices of these vectors focusing on the the first K left-most bits, for different values of K. We created a frequency distribution of such a uniqueness set, as shown on Figure 3. The X axis represents the number of distinct profiles, as counted from the dataset, which correspond to a specific anonymity set (Y axis), ordered from largest to smallest. For example, the point (X=10;Y=1000) indicates that the 10th most popular profile is shared by 1000 users.

In order to improve readability the scale is logarithmic. As is seen on axis X, over $250,000$ of profiles belong to the set 1 (from the axis Y) and thus they are unique. The previous analysis based on subvectors is also presented here. Vectors of sizes 10 (only 10 sites corresponding to maximum of 1024 possible unique profiles) are not enough to create a large set of unique attributes. However, increasing the vector size to 50 provides a very accurate approximation of the data from the full website set. Thus, testing for as few as 50 well-chosen websites in a user's browsing history can be enough to establish a fingerprint which is almost as accurate as when $6,000$ sites are used.

The information-theoretic surprisal is defined as $I = -log_2(F(x))$, where F is a probability density function related to an observation of a particular fingerprint in our dataset. Figure 4 shows a cumulative distribution function of surprisal computed for both Web history and category (vector) profiles. Unique profiles contribute over $18b$ of surprisal and this is the case for the majority of profiles. For general Web history profiles, about 70% of them are close to the maximum surprisal, and for only 50 links in the history this is still the case for about 50% of profiles. Although the maximum surprisal is (unsurprisingly) reached when all sites are considered, it can be approximated by a vector of size 500, as shown in Table 2.

We conclude that the most important sites for fingerprinting are the most popular ones because a considerable number of history profiles are still distinct, even in a small slice of 50 bits. This perhaps counter-intuitive observation may be a result of applying the binomial coefficient: if we assume that in average a user has visited 20 sites of the 50 most popular sites, there are still $\binom{50}{20} = 4.71 \times 10^{13}$ combinations possible. It is worth noting that although users' histories have been tested against more than $6,000$ Web pages, this

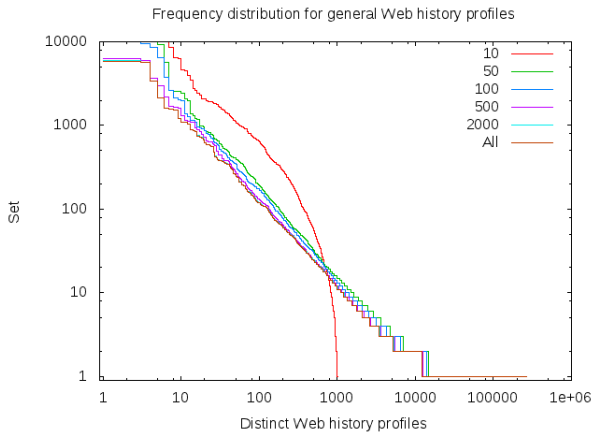


Figure 3: Frequency distributions computed for different bit slices. Even conservatively, only the first $50b$ are sufficient to obtain a large number of unique fingerprints.

does not correspond to a space size of 2^{6000} because of factors such as website popularity, and visitedness correlations based on user interests. Moreover, as is shown in the table 2, it is clear that the dynamics of changes are different than for a uniformly random distribution.

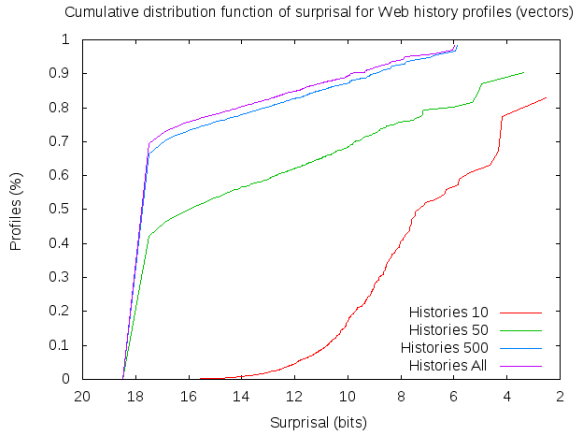


Figure 4: Cumulative distribution function of surprisal for Web history profile vectors of different sizes

The importance of this finding should not be underestimated. If a small number of links is sufficient to prepare an efficient fingerprint for a user, then existing techniques such as DNS or browser cache timing analysis can be used to perform similar analysis.

4.2 Category Profile Uniqueness

4.2.1 Methodology

To extend our analysis, we converted each history profile into a category profile. This was performed by replacing each website of a profile by the general category it belongs to by using the Trend Micro Site Safety Center categorization service [17]. This service takes as input a website URL and outputs a category. Trend Micro uses 72 different interest categories (including news, games, social networking and others).

Category	Count	Category	Count
Search Engines / Portals	1.0000	Social Networking	0.9145
Shopping	0.7891	Computers / Internet	0.7855
News / Media	0.7489	Streaming Media/MP3	0.7229
Entertainment	0.6734	Reference	0.5098
Games	0.2958	Pornography	0.2679
Auctions	0.2436	Government / Legal	0.2385
Software Downloads	0.2201	Blogs / Web Comm.	0.2062
Photo Searches	0.1970	Peer-to-peer	0.1954
Email	0.1666	Business / Economy	0.1552
Sports	0.1250	Financial Services	0.1052
Pers.Net.Stor./File Downl.Srv	0.0950	Travel	0.0920
Adult / Mature Content	0.0588	Education	0.0559
Internet Telephony	0.0521	Chat/Instant Messaging	0.0519
Personals / Dating	0.0516	Internet Radio and TV	0.0493
Vehicles	0.0452	Restaurants / Food	0.0378

Table 1: 30 most popular categories normalized with respect to Search Engines/Portals $(417,750)$. Remaining categories not listed here sum up to 0.35.

4.2.2 Results

We computed a unique set of interests for every Web history profile by discarding repeated occurrences of the same category in profiles. This resulted in $164,043$ distinct category profiles, out of which 88% are unique (i.e. only attributed to a unique user).

Consequently, we can observe that a large number of users have unique personal browsing interests even when analyzed using the more coarse-grained category metric. Figure 1 shows the average number of unique categories per each profile according to history sizes. In a real scenario of an advertising provider, multiple repetitions of each category in the profiles are likely used to enumerate the strength of interest in the category which provides additional information; however, in our analysis, we did not utilize this signal.

Table 1 shows the categories popularity in the profiles, with respect to the most popular category - *Search Engines / Portals*.

Additionally, we converted such category profiles into vectors, similar to the earlier analysis of Web history profiles. The first element of the vector corresponds to the most popular category, i.e. *Search Engines*, the second element to the second most popular category, i.e. *Social Networking*, and so on.

We then computed the frequency distributions for different sizes of the category vector, as in Figure 5.

The associated cumulative distributions are presented on Figure 6. Results show that subvectors of size 30-50 are seen to be enough to prepare a meaningful profile and still maintain a large number of unique profiles. In terms of uniqueness potential, it is sufficient to analyze 30 categories as the data quickly follow the same long tail pattern as in "raw" Web history profiles. However, categories of size 10 are clearly not sufficient because they do not carry enough entropy to distinguish between the large number of profiles in our dataset.

4.2.3 Summary

The conversion from Web history profiles to only use each website's category decreased the overall number of unique profiles. However, we observe that even with the coarser-grained metric there is still a large number of distinct profiles. Assuming conservative profiling and discarding the interest rates for given

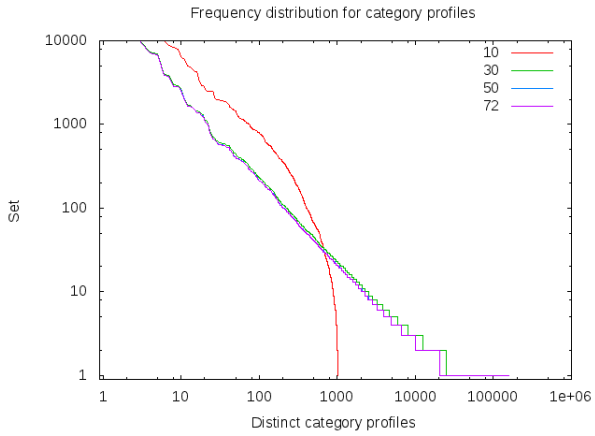


Figure 5: Frequency distributions computed for different bit slices (categories).

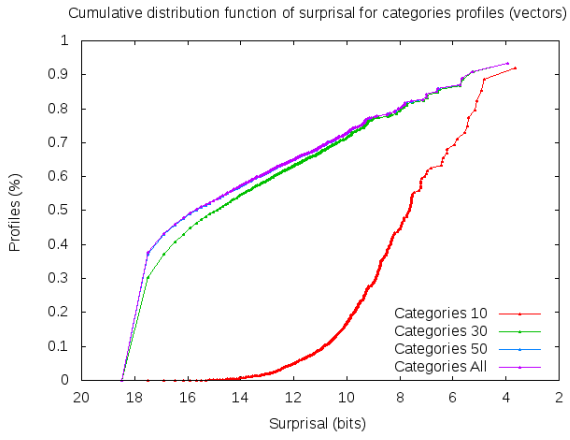


Figure 6: Cumulative distribution (category profiles) function of surprisal for history profiles vectors of different sizes

categories, we were still able to attribute a unique category profile to 39% users.

4.3 Stability of history profiles

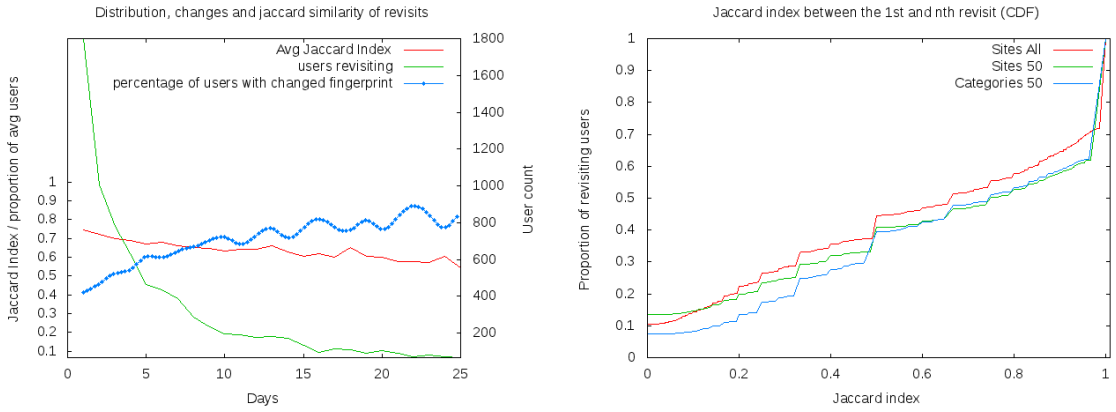
In order to analyze the potential for user tracking using history fingerprints we must understand the stability of Web preferences and browsing history contents. Many Internet users browse the Web on a daily basis, and their browsing history is constantly populated with new websites and subresources. However, all that is not necessarily crucial to address the stability problems since it is sufficient to limit the testing to a small list of sites (e.g. if a user's fingerprint is unique for a subset of n sites, visiting a website outside of the tested set will not affect it).

To quantify this, we analyze history contents of repeat visitors to our test site. Since the dataset is timestamped, it is possible to verify how time affects the history or even if the users cleared their history after learning about websites' ability to detect sites in their history. If a user, identified by the tuple (IP, User-Agent), visited the site on a day_1 and day_n (for $n > 1$), we computed the differences $\{x|day_n - day_1, \text{for all } n > 1\}$ and the similarity between these potentially different Web histories for these two days.

In the analysis only the first 25 days are shown because after this period the number of revisits were small (although some of the users revisited the site even after a year and a half and sometimes the fingerprints for these revisits were also identical). For this analysis we considered profiles constructed from the most popular 50 sites (using subvectors as described before). It can be seen on Figure 7a and it suggests that in considerable number of cases the history remains similar with time, which is especially the case for the first few days after the initial visit. When considering up to 500 most popular sites, the figure do not change significantly.

The cumulative distribution function of the Jaccard Index between the users who re-entered the site is depicted on Figure 7b and as the peak shows, as many as 38% of users had a strongly correlated history.

For 50 subvectors, the fraction of users with a correlated history was 57% and category profile subvectors of size 50 show a similar trend; the number of observed changes is significant. A comparison of this figure to the similar analysis from [5] indicates that Web behavioral fingerprints are slightly less stable than those based on browser configuration for the first few days; the situation becomes comparable for the following days.



(a) **(b)**

Figure 7: (a) The horizontal axis represents the time difference between the first and consecutive visits to the site. The average Jaccard similarity is high for the first days. (b) CDF of Jaccard Index between the fingerprints for the 1st and nth revisit.

Vector Slice	All	Google	Facebook
50	155117	30398	195
100	196899	85682	4130
500	244990	160190	73606
whole	255210	177459	91915

Table 2: Number of unique profiles for different bit slices and vectors: All, Google and Facebook (368284 in total for "popular sites" test)

5 Privacy Risks Analysis

Previously we have shown that browsing preferences are often unique to users and fingerprints derived from them are stable over time in some cases. It is also clear that Web authors can employ known techniques to mount attacks against Web users. Here we focus on a different aspect: what are the possibilities to perform the above analysis by Web service providers? It is important to note that Web service providers have almost constant access to the users' browsing content and are potentially in an ideal position to track the user. In some of the cases users already have accounts on their services which immediately provides information about their identity to the service provider.

5.1 Tracking

Choosing two of the most prominent Web service providers: Google and Facebook and by using real-world scenario data we have verified the extent to which the users may be tracked.

5.1.1 Methodology

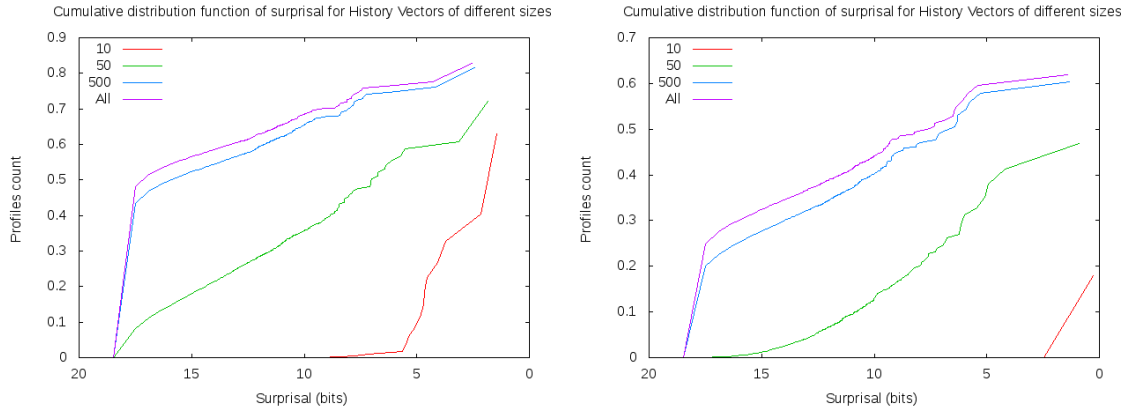
In order to verify to what extent large Web service providers could recreate our profiles, we converted the Web history profiles to the more specific *tracking profiles* which contained only the sites on which we detected scripts from Google and Facebook. In this subsection we also compare them with the data obtained by analyzing "raw" Web histories, as well as category profiles.

5.1.2 Results

After constructing such tracking profiles we computed the distribution of profile sizes and compared them with the one shown on Figure 1. As can be seen, the conversion to tracking profiles decreased the overall sizes of such profiles; however, many of them are still quite large. Figure 8a depicts the distribution of these profiles for "raw" Web history profile (vectors) and these tailored against Facebook and Google services.

We performed the same analysis as in the case of Web history profiles. The numbers of unique fingerprints observed according to the subvector sizes are presented in Table 2. Smaller vector slices for Google and Facebook mean smaller number of unique tracking profiles; for the whole vector ("All") this is sufficient. Smaller vector slices in the former case may result from the nature of the "popular sites" list, which was not prepared to specifically measure this setting. The discrepancy between Google and Facebook is likely due to the fact that Google is providing more Web services than Facebook (AdSense, Analytics, etc.).

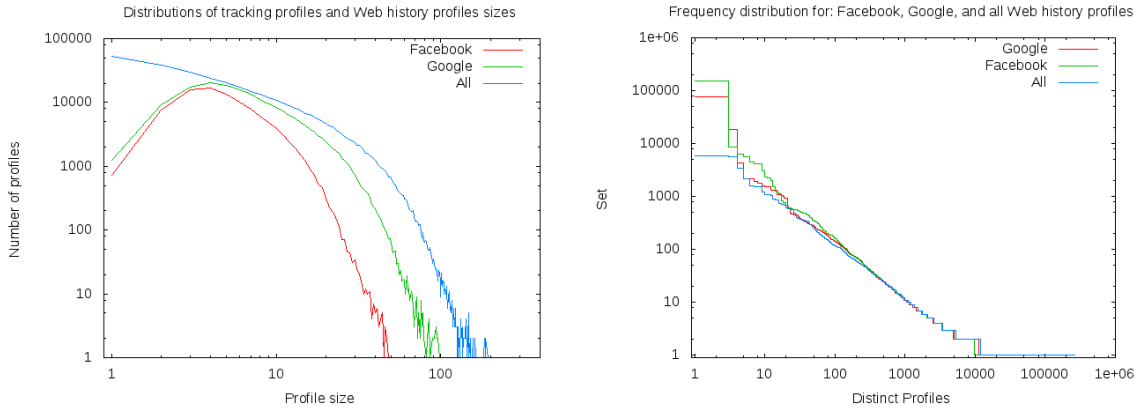
We also computed profile frequency distributions; results are depicted on Figure 8b. The number of unique fingerprints for specific cases of Google and Facebook are smaller, but still considerably large and make analysis possible. They also both occupy larger sets (compared to Web history profiles), and especially in the case of Facebook we observed far less unique fingerprints. The "popular sites" list of pages created



(a) Google **(b)** Facebook

Figure 9: Cumulative distribution (Google and Facebook tracking profiles) function of surprisal for history profiles vectors of different sizes.

for the experiment was not tailored in order to test this particular issue, but as can be seen, there are still many unique profiles which makes this analysis relevant.



(a) Distributions **(b)** Frequency

Figure 8: Distributions and frequency distributions of profiles (All, Google, Facebook).

We then computed the surprisal of the tracking profiles for different bit vector sizes. Figures 9a and 9b show cumulative distribution functions of surprisal for Google and Facebook (vectors). They should be compared to Figure 4. A similar behavior is clearly seen, especially for Google profiles with about 50% of profiles still being unique. The results for Facebook are not as accurate (about 25%).

6 Countermeasures

The problem of information leakage in the modern browsers is not entirely surveyed and researched. Therefore it is not possible to give a perfect and general solution. Particular techniques which allow for history detection can also be merely partially addressed.

The data analyzed in this paper was gathered by performing CSS :visited history detection, which is now generally fixed in the modern browsers, although it will continue to work for older browser installations which constitute to a considerable fraction of Web users. Script blocking may be helpful against certain attacks, but other techniques may still be possible (as was the case with CSS detection which could be conducted without JavaScript). Scripts can, in general, block the 3rd party content present on the sites but this approach often limits the user experience and cannot always be recommended for the average Web user. Plugins (such as AdBlock) blacklisting certain Web resources exist and can defend against this in some cases.

Even after deploying all of available defenses, a user is still prone to other techniques involving timing analysis which can currently only be solved by clearing/removing browser caches, which introduces a usability/privacy trade-off. If a user clears her browsing history and disables browser caching, she will still be prone to other attacks such as DNS cache timing which cannot be cleared as easily.

In the end, the user cannot defend against unknown privacy leaks. If a user wishes to defend against fingerprinting by testing a known pre-defined list of pages, she can ensure that she has visited all of the tested pages—this way, it will be impossible to create a conclusive fingerprint or to enumerate the user's browsing interests. However, we don't expect this to be a realistic mitigation for most Web users.

7 Conclusion

Our work presents the first large-scale analysis of Web browsing histories and their potential for fingerprinting and tracking Web users. In our dataset composed of $368,284$ web histories, more than 69% of users have a unique fingerprint, with a surprisal larger than 18 bits.

The results indicate that Web browsing histories, which can be obtained by a variety of known techniques, may be used to divulge personal preferences and interests to Web authors; as such, browsing history data can be considered similar to a biometric fingerprint. Since Web histories are largely unique to person and, as shown, stable over time in some of the cases, they can be understood as an identifier and potentially be used to strengthening of tracking, in addition to revealing information about a particular user's browsing interests.

We also observed a striking result: it is sufficient to test just a small number of pre-defined sites to obtain vast numbers of different Web history profiles. This indicates that even inefficient techniques for history detection can allow history-based user fingerprinting. Currently there are no known and fully-reliable mitigations to such tracking. By converting visited website data to category profiles, we were able to map the personal interests in much the same way as it is being done by advertising companies. Such profiles, when studying only unique types of categories, were still unique. An analysis of tracking potential (on two examples of Google and Facebook, shown in in Section 5) brings us to a conclusion that Web service providers are also in a position to re-create users' browsing interests.

An interesting question is whether it's possible to extrapolate our data to the entire Internet population. In other words, out of the 2 billion Internet users, what is the percentage of users that have an unique web-history or category profile? As discussed in [5], this number is very difficult, if not impossible, to compute; however, we argue that, as opposed to browser fingerprints, web-histories contain more semantic information that can be exploited to reduce the population size. For example, by considering the language of visited pages, the population size can be reduced from few billions to few millions users. Furthermore, some location-specific sites, such as weather forecast sites, can be used to reduce the population size to few hundred thousands users. Finally, as opposed to the browser fingerprints that rely on a fixed set of browser characteristics, in our case the number of tested or tracked sites can always be increased in order to increase the fingerprint size, and therefore the percentage of unique web-histories.

In the end we believe that Web browsing preferences can be used as an efficient behavioral fingerprint

which is in many cases stable over time. Attacks employing existing techniques such as timing analysis make it simple to conduct large-scale fingerprinting of Web users' interests; this is aided by our observation that only a small set of sites needs to be tested in order to obtain a vast number of unique profiles.

References

[1] Alexa. Alexa 500. `http://alexa.com`.

[2] L. D. Baron. Preventing attacks on a user's history through css :visited selectors. `http://dbaron.org/mozilla/visited-privacy`, 2010.

[3] A. Bortz and D. Boneh. Exposing private information by timing web applications. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 621–628, New York, NY, USA, 2007. ACM.

[4] Bugzilla. Bug 147777 - :visited support allows queries into global history. `https://bugzilla.mozilla.org/show_bug.cgi?id=147777`, 2002.

[5] P. Eckersley. How unique is your web browser? In *Privacy Enhancing Technologies*, pages 1–18, 2010.

[6] E. W. Felten and M. A. Schneider. Timing attacks on web privacy. In *CCS '00: Proceedings of the 7th ACM conference on Computer and communications security*, pages 25–32, New York, NY, USA, 2000. ACM.

[7] C. Jackson, A. Bortz, D. Boneh, and J. C. Mitchell. Protecting browser state from web privacy attacks. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 737–744, New York, NY, USA, 2006. ACM.

[8] A. Janc and L. Olejnik. What the internet knows about you. `http://www.wtikay.com/`.

[9] A. Janc and L. Olejnik. Web browser history detection as a real-world privacy threat. In *ESORICS*, pages 215–231, 2010.

[10] D. Jang, R. Jhala, S. Lerner, and H. Shacham. An empirical study of privacy-violating information flows in JavaScript Web applications. In A. Keromytis and V. Shmatikov, editors, *Proceedings of CCS 2010*, pages 270–83. ACM Press, Oct. 2010.

[11] B. Krishnamurthy and C. Wills. Privacy diffusion on the web: a longitudinal perspective. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 541–550, New York, NY, USA, 2009. ACM.

[12] S. Krishnan and F. Monrose. Dns prefetching and its privacy implications: when good things go bad. In *Proceedings of the 3rd USENIX conference on Large-scale exploits and emergent threats: botnets, spyware, worms, and more*, LEET'10, pages 10–10, Berkeley, CA, USA, 2010. USENIX Association.

[13] B. Miller. Vital signs of identity. *IEEE Spectr.*, 31:22–30, February 1994.

[14] R. Moskovitch, C. Feher, A. Messerman, N. Kirschnick, T. Mustafic, A. Camtepe, B. Löhlein, U. Heister, S. Möller, L. Rokach, and Y. Elovici. Identity theft, computers and behavioral biometrics. In *Proceedings of the 2009 IEEE international conference on Intelligence and security informatics*, ISI'09, pages 155–160, Piscataway, NJ, USA, 2009. IEEE Press.

[15] K. Mowery, D. Bogenreif, S. Yilek, and H. Shacham. Fingerprinting information in javascript implementations. In *Proceedings of W2SP 2011*, IEEE Computer Society, 2011.

[16] Quantcast. Quantcast. `http://www.quantcast.com/`.

[17] Trendmicro. Trend micro site safety center. `http://global.sitesafety.trendmicro.com`.

[18] W3Schools. W3schools online web tutorials. `www.w3schools.com`.

[19] Z. Weinberg, E. Y. Chen, P. R. Jayaraman, and C. Jackson. I still know what you visited last summer: Leaking browsing history via user interaction and side channel attacks. In *Proceedings of the 2011 IEEE Symposium on Security and Privacy*, SP '11, pages 147–161, Washington, DC, USA, 2011. IEEE Computer Society.

[20] C. E. Wills, M. Mikhailov, and H. Shang. Inferring relative popularity of internet applications by actively querying dns caches. In *Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement*, IMC '03, pages 78–90, New York, NY, USA, 2003. ACM.

[21] G. Wondracek, T. Holz, E. Kirda, and C. Kruegel. A practical attack to de-anonymize social network users, ieee security and privacy. In *IEEE Security and Privacy*, Oakland, CA, USA, 2010.

[22] R. V. Yampolskiy and V. Govindaraju. Behavioural biometrics: a survey and classification. *Int. J. Biometrics*, 1:81–113, June 2008.

[23] T.-F. Yen, X. Huang, F. Monrose, and M. K. Reiter. Browser fingerprinting from coarse traffic summaries: Techniques and implications. In *Proceedings of the 6th International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, DIMVA '09, pages 157–175, Berlin, Heidelberg, 2009. Springer-Verlag.

[24] T.-F. Yen, Y. Xie, F. Yu, R. P. Yu, and M. Abadi. Host fingerprinting and tracking on the web:privacy and security implications. In *19th Annual Network and Distributed System Security Symposium (NDSS) 2012, Internet Society*, 2012.

[25] M. Zalewski. Browser security handbook, part 2. `http://code.google.com/p/browsersec/wiki/Part2`, 2009.