Browsing Unicity: On the Limits of Anonymizing Web Tracking Data

Clemens Deußer
Chair of Privacy and Security
TU Dresden, Germany
Email: clemens.deusser@tu-dresden.de

Steffen Passmann
INFOnline GmbH
Berlin, Germany
Email: SPassmann@infonline.de

Thorsten Strufe
Karlsruhe Institute of Technology
Centre for Tactile Internet, TU Dresden
Email: strufe@kit.edu

***Abstract*—Cross domain tracking has become the rule, rather than the exception, and scripts that collect behavioral data from visitors across sites have become ubiquitous on the Web. The collections form comprehensive profiles of browsing patterns and contain personal, sensitive information. This data can easily be linked back to the tracked individuals, most of whom are likely unaware of this information's mere existence, let alone its perpetual storage and processing. As public pressure has increased, tracking companies like Google, Facebook, or Baidu now claim to anonymize their datasets, thus limiting or eliminating the possibility of linking it back to data subjects.**

In cooperation with Europe's largest audience measurement association we use access to a comprehensive tracking dataset to assess both identifiability and the possibility of convincingly anonymizing browsing data. Our results show that anonymization through generalization does not sufficiently protect anonymity. Reducing unicity of browsing data to negligible levels would necessitate removal of all client and web domain information as well as click timings. In tangible adversary scenarios, supposedly anonymized datasets are highly vulnerable to dataset enrichment and shoulder surfing adversaries, with almost half of all browsing sessions being identified by just two observations. We conclude that while it may be possible to store single coarsened clicks anonymously, any collection of higher complexity will contain large amounts of pseudonymous data.

I. Introduction

Tracking has become pervasive on the Web. More than four out of five sites employ behavioral tracking, some on a large scale, with dozens of different scripts tracking their users at the same time [1], [2]. The average page access on the Web is tracked by eight scripts, today[1]. Some sites employ local tracking to optimize their user experience, others use legitimate scripts to perform reliable audience and reach measurements. The majority of trackers, however, is used to presumably improve targeted advertisement [3], [4].

While the request to clear one's browser history in case of emergency has made it into contemporary folklore due to how sensitive such data is, a broad industry has been establishing increasingly comprehensive overviews of browsing histories of users across essentially the entire Web. Upon visits to Web pages, tracking scripts identify the browser across websites and store entire browsing profiles or sequences of observed visits as so called click traces in vast tracking databases [5].

The usual reflex to inquiry is the statement that this data was anonymized, usually through generalization (truncation, or "coarsening") of stored attributes, such as IP addresses [6] or through differential privacy techniques. Differential privacy is a powerful tool which delivers provable privacy guarantees. In this paper we will not examine practical implementations of differential privacy, but in the past they have often been either misused (eg. through lack of a properly enforced privacy budget) or have lead to severely restricted utility [7], [8]. Instead we will focus on examining generalization techniques.

Whether and how generalized data can be de-anonymized has been extensively researched by Narayanan et al. and others in the past [9], [10], [11]. Nevertheless, anonymization through generalization techniques not only continue to be used, but the industry in which they are applied plays an increasingly ubiquitous role in modern society. Their position is that these results are not universally valid and do not apply to other methods of generalization on different kinds of data. In this work we will attempt to close that gap as it relates to web tracking data. More specifically, both structural information, such as position in a social graph, as well as pseudonyms in general have been shown to be highly identifying. In recognition of this fact, modern privacy regulations like the European GDPR specifically enforce restrictions such as obtaining informed consent before allowing collection and processing of pseudonymous data. Storing a client browsing session as a sequence of website visits with very general page and client information, as audience measurement providers often do, appears to avoid these restrictions.

We argue that a combination of attributes and sequential information can be uniquely identifying as well and thus constitutes an implicit pseudonym. Once enough elements of a browsing sequence have been observed, the entire session can be linked back to the data subject. Shoulder surfing - physically or through digital dossier aggregations - is one example where this fact can be exploited. Another is the trading of supposedly anonymized data between tracking companies, where the buyer can match unique partial traces to their own data. In that way they gain access to the browsing history of data subjects they did not track themselves, evading the obtaining of consent and data protection rules.

This work has in parts been supported by the German Research Foundation DFG, the Cluster of Excellence EXC 2050/1 "Centre for Tactile Internet" (CeTI) as part of Germany's Excellence Strategy, and INFOnline GmbH.

[1]https://www.whotracks.me

DOI 10.1109/SP40000.2020.00018

In this paper we will not only attempt to show how easily such pseudonyms can manifest, but we will also investigate whether the techniques applied by industry can prevent the emergence of pseudonyms in tracking databases at all. More specifically, we ask ourselves the following two questions: (1) How frequently do pseudonyms emerge in anonymized tracking data? and (2) How easily can tracking data be linked to secondary sources? In this course we aim to understand to which extent coarsening the available data can actually help to reduce identifiability.

We analyze an obfuscated sample of the data of the largest technical provider for German Audit Bureaus of Circulation, one of Europe's largest providers for audience measurement services. The analyzed sample contains 65.2 million clients and over 2.3 billion page impressions. We adhere to industry standards for data treatment to generate a database of click traces. Following the rationale of IP address truncation, we then successively reduce click trace length as well as the level of detail of the available information per click; including information about the visited page, the timestamp, and data collected from the browser.

We then calculate the unicity, the fraction of unique click traces, as a measure of how pseudonymous the data is. We argue that a unique browsing session is by itself a pseudonym and thus cannot be anonymous. To test anonymity in a more practical vein we also act as an adversary in the two scenarios mentioned above - shoulder surfing and data exchange enrichment.

As we have already acknowledged, the data we analyze is highly sensitive and private. We take the responsibility of working with such data very seriously. The data was accessed solely through scripts run locally on the database servers, directly generating the results we present here. As a consequence we do not have direct access to said data and cannot provide it. Verification of our results can still be facilitated through the same method we used, at the discretion of the database owner.

II. Background

Web browsing behavior is processed for different reasons. To optimize the browsing experience on their website, web developers have long parsed web server access log files, and later turned to entirely *local, per domain tracking* scripts, like Matomo (formerly Piwik)[2].

Site analytics and cross domain tracking (Google, Facebook, Yandex Metrica, etc.) provide web developers with similar functionality. Their business model is based on collecting browsing behavior across several sites, mainly to improve advertisement accuracy and allow extended features like retargeting. Their reach varies. Some have managed to be present in the vast majority of the popular Web [2]. Smaller tracking companies started to extend their data by trading tracking data with competitors. User data exchanges provide such markets, buyers can even bid for the profiles of specific users [12]. This market has grown to dozens of providers. So meta tools have emerged that manage the combination of trackers that are used for specific page calls, based on chosen policies[3].

Another reason is *audience measurement* [13]. The market of publishers and advertisers on the Web requires independent third parties. These ABCs (Audit Bureau of Circulation) verify the popularity of sites, and their claimed number of visits. They measure the performance of advertising media, to provide the advertisement market with indicators of the relevance of the different outlets.

A. Internet Audience Measurement

Audience measurement is traditionally conducted with panels or full evaluations. For this paper, we focus on the latter, as it implements a census measurement similar to Web tracking. Technically, this is implemented by injecting JavaScript snippets ("tags") into the code of a Web page. It collects and sends information to the tracker when the page is rendered on the client system.

The requirement for cross-market data leads to the implementation of a third-party approach. Hence, both third-party JavaScript and third-party cookies are used[4]. When integrating the script into the website, the publisher provides two essential pieces of information in the html-tag. An identifier for the website as a whole (website-identifier) and an identifier for the specific, visited page, called "code".

The transmitted dataset is received by a web server of the measuring system. It is then enriched and stored as a tuple of the client ID (extracted from a cookie), the geolocation of the client (as queried based on its IP address, to an accuracy of the federal state level), and a timestamp. Similarly, the user agent is converted to an estimated "device type" using a corresponding database.

The publisher provides further information. This includes the categorization of all pages contained in their site, according to the standards of the International Federation of ABCs (IFABC).For each unique combination of website-identifier and code, the publisher provides a number of features. These include the category of the content (news, social, sports, politics[5]), the media (image and text, video), the creator (editorial content, user-provided), the language, whether it is paid content, whether it is the entry-page (i.e. index.html) and for which device the exact page was optimized (desktop, mobile).

ABCs then publish the essential results of activity on the measured sites, which is usually the number of page impressions, visits, and clients. The IFABC defines a *page impression* as "[...] every user-induced action (e.g. a click) that leads to a significant change in the view [...]". This definition includes scrolling pages with progressive loading. *Visits* are defined as

[2]https://matomo.org

[3]https://marketingplatform.google.com/about/tag-manager/

[4]Browser developers have recently started to prevent 3rd-party cookies, so many large trackers now turn to integrating 1st party content to the pages, thus being able to also set 1st-party cookies, or to exploit session resumption of TLS [14].

[5]https://support.aerserv.com/hc/en-us/articles/207148516-List-of-IAB-Categories

sessions of consecutive page impressions with an inter-arrival time of 1800 seconds (30 minutes) or less. *Clients* represent unique, returning visitors. Several visits can correspond to the same client, and since client IDs change (e.g. when cookies are deleted or various devices used), a multitude of measured clients can correspond to the same individual.

Tracking databases essentially contain sequences of action entries. Each entry traditionally consists of extensive client and page information, such as IP address, unique ID (cookie), user agent, visited URL, page category (topical), timestamp and many more[6]. IP addresses nowadays are truncated as privacy regulations prohibit processing of explicit identifying information without consent.

B. Pseudonymity and Threats

Browsing data is highly sensitive. This is especially true for cross domain tracking: the same trackers from very few, large companies are found in the majority of websites offering medical advice, information on planned parenthood, opinion formation, political discussion, even pornographic content, and also web search and social networking [2]. Activities across these sites are linked by their client ID to sessions in the tracking databases. Some entries may contain plaintext pseudonyms or even names as parameters of stored URLs.

Tracking companies contest such concerns, maintaining that they do not attempt to identify individuals, and that they anonymize their databases. However, even when measures such as IP address truncation and removal of URL parameters and other directly identifying information are correctly and faithfully applied, they may not actually anonymize the data. This is because the stored data pertaining to an individual remains pseudonymous as long as the connection to the data subject is unique. Meaning there is no other individual exhibiting the exact same data signature. As long as this pseudonym exists, it can in principle be linked back to the individual identity. While we will present tangible scenarios as to how this can happen even today, it is clear that more sophisticated techniques and more massive databases will be available in the future and potentially retroactively expose pseudonyms in today's databases. Privacy regulations therefore impose severe restrictions on the use of data that is not strictly anonymous. The GDPR for instance inversely defines "anonymous" as "the data subject is no longer identifiable". It thus requires that data subjects can no longer be linked against the data, which also precludes the existence of pseudonyms.

In this paper, we follow the interpretation of the GDPR and consider any information that implicitly identifies an individual a pseudonym. The stored client ID that ties the clicks of sessions together of course represents such a pseudonym. However, being assigned randomly, it may not be easy to link it back to individuals. The behavior as encoded in the click trace on the other hand may also represent a pseudonym. This clearly holds for all unique click traces, which therefore have to be considered pseudonyms in themselves. Furthermore, we expect that external information exists in abundance that facilitates linking the click trace to an identity.

To emphasize possible *threats*, we describe two commonplace scenarios, which provide trackers with the possibility to identify the users behind click traces in their databases, given that some of them are unique.

We first consider a tracking company that partakes in a *user data exchange*, like BDEX, BIG, onAudience, or Lotame. Most websites employ a combination of trackers, and some user data inevitably ends up in databases of different trackers, causing an overlap of page views between trackers. This overlap can be used to identify click traces in acquired data, to enrich the owned database with additional profiles. This also allows for the re-identification of click traces in acquired data, thus learning additional, potentially sensitive, and explicitly identified activities.

Second, we consider different types of *shoulder surfing*. Assuming the tracking database contained pseudonymous click traces that are unique in less than the entirety of their clicks, partial knowledge could suffice to identify individuals that are represented in the data. This requires observation of an identifying subset of clicks, which is easy to imagine: The textbook example is a colleague or bystander who watches another user. Considering frequent public sharing of links on social media, a much more scalable and globally available way to collect such identifying sets of clicks is to automatically scrape social media sites, and filter the posts of real-name profiles for shared links.

Tracking companies argue they are not interested in identifying individuals. The data of a large fraction of them, however, is readily available for low prices at user data exchanges, and data loss incidents happen to even the largest and most profitable companies[7,8].

To evaluate the occurrence of pseudonyms in tracking databases, we will analyze a representative, real dataset.

III. Identification metrics and anonymization strategies

In this section, we will introduce the metrics we use to measure pseudonymity and identifiability. We subsequently discuss strategies that commonly are suggested to anonymize tracking databases.

The data collected by trackers upon a page call is commonly stored as a tuple of page and client information, which we call a *click*. It is possible to assemble the set of clicks of a client by selecting the tuples with identical client identifiers. Depending on browser settings, these client identifiers may change frequently, resulting in very small sets of clicks, or remain stable over long periods of time resulting in very large collections of clicks of the same client. To obtain a more consistent dataset, we do not consider these client sets

[6]Overviews are at https://developer.matomo.org/api-reference/tracking-api or https://developers.google.com/analytics/devguides/collection/protocol/v1/parameters.

[7]https://www.theverge.com/2018/10/8/17951914/google-plus-data-breach-exposed-user-profile-information-privacy-not-disclosed

[8]https://www.wired.com/story/facebook-security-breach-50-million-accounts/

and instead only retain information pertaining to individual browsing sessions. We call the sequence of clicks representing a browsing session a *click trace*.

Definition 1 (Clicks and click traces): A *click* is a tuple representing a page impression. It may contain information about the page, a client identifier, other characteristic data about the client as well as a timestamp. Consider a database of browsing data consisting of individual clicks. A *click trace* is an ordered sequence of clicks. All clicks of a click trace belong to the same user and are ordered chronologically.
A click trace β is a *subsequence* of click trace α, or $\beta \subseteq \alpha$, if all its clicks are contained in α in the same order (not necessarily subsequently, so $AC \subseteq ABC$, but $BA \not\subseteq ABC$). A *traceset* is an unordered set of click traces. The n-*subset* of a click trace α is the collection of all its subsequences of length n, or more formally: $N_n(\alpha) = \{\beta \mid \beta \subseteq \alpha, |\beta| = n\}$ with $|N_n(\alpha)| = \binom{|\alpha|}{n}$.

In this paper we want to assess how easily pseudonyms emerge, and tracking data can be identified using secondary sources.

A pseudonym can be seen as the lowest form of identity. It does not directly identify the subject, but it uniquely corresponds to one. Once that link is found out, for example through some external data, the pseudonymity is broken and the data subject is identified. By measuring the degree to which pseudonyms exist in a collection of click traces, we can determine how vulnerable it is to simple de-anonymization.

Applied to a database of click traces, our definition of pseudonyms relates to k-anonymity. A database contains no pseudonyms if it is k-anonymous with $k \geq 2$; in other words, if no click trace is unique. However, the binary nature of k-anonymity severely limits its utility for our purpose. It is to be expected that at least one click trace will remain unique under most coarsening measures, for example because a web page may only have a single visitor or a location only a single browser. k-anonymity would never be fulfilled, even if coarsening measures were relatively successful. Instead, the fraction of unique click traces allows for a more nuanced observation. For that reason we adopt the approach used by De Montjoye et al. [15], which defines unicity as the proportion of unique pieces of information. Unicity in this case serves as a measure of how close the database is to being anonymous. A unicity of 0 implies k-anonymity, a unicity of 0.5 means half the click traces in our traceset are pseudonyms.

Definition 2 (Unicity): We say that click traces α and β are equal, or $\alpha = \beta$, if and only if $\alpha \subseteq \beta$ and $\beta \subseteq \alpha$. If and only if two click traces are equal they belong to the same *anonymity set*. A click trace is *unique* if it is the only member of its anonymity set. The *unicity* of a traceset is its ratio of unique click traces over all click traces.

As previously mentioned it is not always necessary to know all clicks from a trace to make it uniquely identifiable. While the pseudonymity of a unique click trace is by itself first and foremost a theoretical issue, unique partial traces present an immediate practical adversary model. Note that Partial information is more easily obtained and, once identified, the click trace contains previously unknown information. We want to find out how little information is necessary for successful identification and call the corresponding metric *identifiability*. The idea behind that metric is as follows: given an adversary with some well defined capability of obtaining partial information of some browsing session, an identifiability of 0.2 means that the corresponding full click trace has a 20% chance to be identified. It is important to note that the adversary in this model does not actually have the partial information, in which case the corresponding click trace would either be or not be identified. Rather he has the abstract capability, represented by a set of possible partial traces he can draw from, and identifiability is then the share of samples in this set uniquely identifying the original trace.

Definition 3 (identifiability): The *compatibility class* $\theta(\beta, T)$ of click trace β given traceset T consists of all click traces $\alpha \in T$ such that $\beta \subseteq \alpha$. We say that a click trace $\alpha \in T$ is *identified* by β, or β identifies α, if α is the only member of its compability class, or $\theta(\beta, T) = \alpha$. Given traceset I_α, the *identifiability* $\rho_\alpha(T, I_\alpha)$ of click trace $\alpha \in T$ is the ratio of click traces $\beta \in I_\alpha$ that α is identified by.

The weighted identifiability of a trace set T given $I = \{I_\alpha | \alpha \in T\}$ is

$$\rho(T, I) = \frac{\sum_{\alpha \in T}(|\alpha| \rho_\alpha(T, I_\alpha))}{\sum_{\beta \in T} |\beta|}$$

I_α represents the adversary, or rather all the possible ways with which they might attempt to identify α. For example, I_α might consist of all subtraces $\beta \subseteq \alpha$ of length $|\beta| = n \leq |\alpha|$, or $I_\alpha = N_n(\alpha)$, representing an adversary making exactly n random observations of click trace α.

Success of the adversary measured in the identification ratio depends on the adversary's prior knowledge. An adversary knowing the entire dataset can clearly identify every pseudonym, but gains no information in the process. An adversary knowing very little may identify only a few click traces, but learn much more in the process, relatively speaking.

We now describe the implementation of click trace generation, as well as the calculation of unicity and identifiability.

A. Extracting Click Traces

The dataset which we are going to use for our evaluation does not contain click traces, rather it contains the full browsing history of each client, including a unique ID. As a first data processing step we thus need to turn the full browsing history into individual browsing sessions.

The database we analyze is extremely comprehensive, we made some implementation decisions to achieve feasible calculation. Due to technical considerations the database can only be accessed sequentially. We can choose an order in which entries are processed a priori, but we cannot access them out of order. Following the industry definition of a browsing session (ref. to section II), we build click traces iteratively by pushing clicks from a chronological click stream until two consecutive clicks are more than 1800 seconds apart or the trace exceeds a given maximum length (Algorithm 1).

input : chronologically sorted stream C, max length ml;
all $c \in C$ contain timestamp c_t and click trace ID c_i
output: traceset T
$T \leftarrow \{\}$; TempTraces $\leftarrow \{\}$; LastTime $\leftarrow \{\}$;
for $c \in C$ **do**
  if $c_i \in$ *TempTraces* **and** $c_t - LastTime[c_i] < 1800$ **and** *TempTraces*$[c_i] < ml$ **then**
    TempTraces$[c_i] \leftarrow$ TempTraces$[c_i] \cup c$;
  else
    $T \leftarrow T \cup$ TempTraces$[c_i]$;
    TempTraces$[c_i] \leftarrow c$;
  end
  LastTime$[c_i] \leftarrow c_t$;
end
for *trace* $\in$ *TempTraces* **do**
  $T \leftarrow T \cup$ trace;
end

Algorithm 1: Calculating click traces from data stream

B. Calculating Unicity

Computing unicity requires a number of comparisons between click traces to determine whether they belong to the same anonymity set. By hashing click traces and using an index we only require logarithmic time to determine whether a click trace is unique or not. The overall complexity of Algorithm 2 is thus $O(n\ log\ n)$.

input : traceset T, click trace properties w, hash function h
output: unicity and anonymity sets *Anon* of T
Anon $\leftarrow \{\}$
for $w_i \in w$ **do**
  for $t \in T(w_i)$ **do**
    /* check if t's anonymity set already exists*/
    if $t \in$ *Anon* **then**
      Anon(t) $\leftarrow$ Anon(t) $+1$;
    else
      Anon(t) $\leftarrow 1$;
    end
  end
end
unique $\leftarrow 0$;
for $t \in$ *Anon* **do**
  if *Anon*$(t) = 1$ **then**
    unique $\leftarrow$ unique $+ 1$;
  end
end
unicity $\leftarrow \frac{\text{unique}}{|T|}$

Algorithm 2: Unicity and anonymity sets given a traceset

We can further reduce the computation time by grouping click traces by their coarsened timestamp range and length. Doing so does not reduce time complexity, but significantly increases performance.

C. Evaluating Identifiability

Algorithm 3 calculates identifiability according to definition 3. We assess identifiability with two realistic threats in mind: (1) the case of database trading, and (2) shoulder surfing. Identifiability depends on prior knowledge of the adversary, which differs between these scenarios and is represented in the adversary set.

For *web trackers* we define prior knowledge as the fraction of the acquired dataset overlapping with their own data.

Recalling definition 3, for each adversary we need to define the adversary set I. For an acquired dataset T and a given overlap $\sigma \in [0, 1]$, we consider all possible sets of websites W such that the fraction of clicks belonging to each such set is within an ϵ of the overlap: $\forall \omega \in W, \left| \frac{\sum_{\alpha \in T} |\gamma_{\alpha,\omega}|}{\sum_{\alpha \in T} |\alpha|} - \sigma \right| < \epsilon$, with $\gamma_{\alpha,\omega}$ being the maximal subtrace of α such that each click belongs to a website in ω. Analogous to definition 3 we obtain the adversary set

$$I = \{I_{\alpha,W} | \alpha \in T\} \text{ with } I_{\alpha,W} = \{\gamma_{\alpha,\omega} | \omega \in W\}.$$

For example with an overlap of 0.2 an identifiability of 0.5 means that an adversary, who acquired an external dataset with 20% of its clicks appearing in his own data, can uniquely identify half of the click traces in the acquired dataset. Having re-identified click traces in the acquired data, the adversary learns additional actions of the client on sites that he does not track. We define *gain* accordingly to be the fraction of clicks of the acquired dataset belonging to identified click traces, which are not contained in the known dataset.

For the *shoulder surfer*, we define prior knowledge as the number n of observations known to the adversary, therefore

$$I = \{I_\alpha | \alpha \in T\} \text{ with } I_\alpha = N_n(\alpha).$$

input : acquired traceset T, adversary set I, sample size s
output: identifiability *ident* of T
index $\leftarrow 0$;
for $t \in T$ **do**
  /* build associative array assigning clicks to their trace */
  traces[index, index $+ |t|$] $\leftarrow t$;
  index $\leftarrow$ index $+ |t|$;
end
IndexSamples $\leftarrow$ draw s samples from $[1, \text{size(T)}]$;
count $\leftarrow 0$;
for i *in IndexSamples* **do**
  $\alpha \leftarrow$ traces(i);
  subsample $\leftarrow$ draw sample from I_α;
  matched $\leftarrow$ False;
  for $\beta \in T$ **do**
    if *subsample* $\subseteq \beta$ **then**
      matched $\leftarrow$ True;
    end
  end
  if *not matched* **then**
    count $\leftarrow$ count $+1$;
  end
end
ident $\leftarrow \frac{\text{count}}{s}$;

Algorithm 3: Calculating identifiability from a traceset

Identifiability cannot be calculated using hashed values the way we calculate unicity, as we have to determine whether a smaller click trace is contained within a larger one rather than whether they are equal. In addition, the adversary set I is far too large to allow exact calculation. For example, calculating identifiability for a shoulder surfer making 3 observations on a million click traces of length 10 requires a number of

operations on the order of $14.4 \cdot 10^{15}$. Our database contains far more than a million click traces, some several thousand clicks long. It is clear that exact identifiability cannot be computed. Instead, we follow the approach of De Montjoye [15] and approximate the true value using established sampling techniques.

In order to approximate identifiability by sampling, we need to sample from the set of all possible results. The set of all possible results, again, is far too big to be computed. Instead, for algorithm 3 we sample results by selecting the click trace of a random *click* from all click traces (thus selecting a click trace weighted by its length), and then selecting from all possible attack configurations, given the adversary's capabilities.

Sampling in this way corresponds to a series of Bernoulli trials, which allows us to use established formula for sample size n_0 given a confidence interval and error estimation.

$$n_0 = \frac{Z^2 p(1-p)}{e^2}$$

This expression is maximized for $p = 0.5$, which is our best estimation since p is unknown. For a confidence of 99% ($Z = 2.576$) and a maximum error of 1% ($e = 0.01$), meaning that any subsequent experiment has a 99% chance of deviating from the result by at most 1%, we obtain a necessary sample size of $n_0 = 16,590$. The expected necessary granularity of our results is well below 1%, so we can use this sampling size for all our identifiability experiments.

D. Anonymization

Understanding tracking databases and ways to identify users that have generated the contained traces, we turn to strategies that are commonly suggested for their anonymization.

Considering the composition of entries in tracking databases, as described in Section II-A, we group the parameters into (1) information about the user (IP, client ID, user agent, location), (2) information about the visited page (URL, category), and (3) information about the access (method, referrer, timestamp).

Since privacy regulations require either informed consent or the absence of identifying information, trackers have confined themselves to store only truncated IP addresses. Generalizing the direct identifier, they claim that the processed data thus was anonymous.

It is easy to see that the data above still contains pseudonyms. Storing a page call with its exact time in milliseconds creates a unique identifier with high probability, as exactly simultaneous calls to the same page are unlikely on that time scale.

We hence explore anonymization of the described groups of parameters, following the same vein of generalization. The most intuitive measure is to coarsen the *timestamps*. We do so by removing the least significant time information, similar to truncating bits of the IP address. Specifically, we coarsen a timestamp by subtracting the timestamp modulo the coarseness parameter. For example a timestamp of 152.9867 with a coarseness parameter of 60 seconds is coarsened to 120. We use coarsening parameters up to the order of $100,000$ seconds, which is slightly over a day and sufficiently below the scope of the analyzed data (ref. section IV).

The *visited page* and all *information about the user* can also successively be discarded, to reduce unicity in the click traces. Some properties are suited for gradual anonymization: the page can be generalized from the exact URL to the code of the page, its category, or simply the FQDN of the visited website. The same holds for information about the user, where we can remove information about the user-agent and geolocation.

Finally, some properties of the traces directly relate to unicity: traces collected across several websites contain more information than click traces that are restricted to single websites only. The length of the clicks that are linkable to a single session also correlates to identifiability, as long traces contain more information and are much more likely to be unique than short traces. Restricting the maximum length of click traces, or limiting them to single websites, are other possible strategies when aiming to anonymize datasets.

IV. Data

For this study we joined forces with the audience measurement provider of an ABC representing a majority of German websites. It spans over 2500 websites and apps in total, with an average volume of 2 to 3 billion page impressions per day. This data is thus representative for the German market, but we cannot say with certainty whether our experiments would yield the same results using data of another provider, such as Google Analytics. Nationality likely does not have an effect as we are analyzing meta data rather than content, but to our best knowledge this specific subject has not yet been explored in literature.

The ABC stores this data for the purpose of calculating quantitative session metrics, like visits and returning clients. It stores a subset of common tracking parameters, as described in Table I.

First, each entry contains a client ID, tied to a session cookie. In our experiments we use this ID only to assemble the database of click traces, but discard it before assessing unicity and identifiability. A geolocation is stored on the granularity of federal states, determined by looking up the IP address of the browser in a public database, and the IP address is subsequently discarded. The ABC also stores a page *code* of the visit. This code is assigned by the publisher, and usually encodes an article, or specific site (local path of the URL), as well as the user-agent for which its layout has been optimized. Additional information about the visited page are the *site* and its *category*. The former corresponds to the public host part of the URL, or FQDN, and the latter to classes of content, as they are defined by the ABC (sports, politics, etc). Finally, each entry contains the time of the user's click, stored as a Unix timestamp with millisecond precision.

Some of the page-related information has global, and some local characteristics. Categories on the one hand are global to the ABC, so different sites will have pages with identical categories. The codes, on the other hand, are chosen by

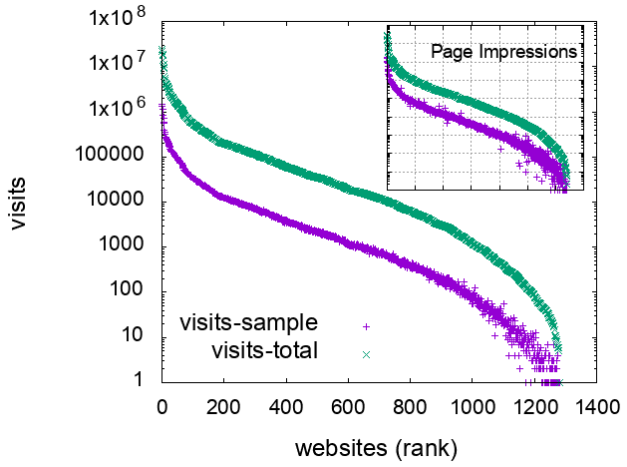

Fig. 1. Distribution of visits: sample vs entire dataset (PIs in inset). Websites are ranked by the number of visits in the original dataset. The sample distribution largely follows the original distribution, meaning websites are represented proportionally in the sample.

the respective publishers for their own site. They only have significance for their respective site and may even overlap with codes of other sides. Therefore the code information can only be used if the site information is used as well.

Note that the only explicit information beyond the ID that is stored about the clients is their geolocation. The choice of device-type and browser may be implicitly represented in the page code.

Field	Content
Timestamp	Unix timestamp in microseconds
Client ID	Unique per user / browser, from cookie
Site	ID of visited website/FQDN
Code	ID of displayed page, assigned by publisher
Category	Category of page, according to ABC
Geolocation	DB lookup of client IP

TABLE I
INFORMATION STORED PER CLIENT ACTION

The entire database of the measurement provider is far too large to analyze unicity and identifiability. Behavior on the Web being driven by freshly published content, we analyzed an interval of one week in March 2019. We limited our dataset to desktop clients that accept cookies and do not exhibit any characteristic behavior (for example search engines, bots, etc). This produced a highly reliable and clean dataset without requiring additional preprocessing. As a result, we do not take into account mobile browsing and we acknowledge that as a limitation of this study. The full week's worth of data contains 66.1 million clients, 2.34 billion page impressions, and 351.3 million visits.

From this data we sampled a 16th of the clients at random from roughly half of available sites, to reduce the sample to a size that allowed computation of unicity given the resources available. To verify that our results are representative, we ran all experiments on samples of increasing size, and empirically observed that the experiment outcome converged with increasing sample size well before reaching the size of our final sample.

To validate our sample, we performed some basic sanity tests: Fig. 1 compares the frequencies of page impressions and visits of the original data vs. the sample. We observe that the frequency distributions in the full census and our sample follow equal characteristics and we verified that our click trace lengths comply to literature [16].

The final sample contains data as described in Table II.

PIs	Visits	Clients	Locations	Sites	Codes	Categories
147.9M	22.1M	4.1M	3053	1281	62.5K	725

TABLE II
COMPOSITION OF THE TESTED SAMPLE

While analyzing large data vaults is always a challenge, the resource constraints we experienced would not necessarily apply to an adversary. They may only be interested in a single click trace while we analyze large samples in a multitude of different scenarios. In addition, they may have access to resources far exceeding ours.

Note that at no point in our experiments was plain text data analyzed. All the adversary models are only applied on a theoretical level, meaning no actual user was de-anonymized. Or, in other words, no privacy was harmed in the making of this study. For regulatory reasons we cannot share the final dataset, but we will provide access to run reproduction studies upon request.

V. EMPIRICAL RESULTS

Our main interest in this paper is to assess to which extent pseudonyms emerge in tracking databases. We divide this general question into two studies over increasingly generalized data, investigating first the unicity of the data and afterwards the identifiability. In the following section we describe how we conducted our experiments and report the results.

A. Experimental Setup

The experiments of this paper were computed on a small standard hadoop platform with about $2,000$ cores. All experiments were classical map-reduce jobs. A proven mapper was used for all applications. The reducers were developed according to the requirements of the respective experiment.

Within the experiments, we are searching for unique click traces. This terminally requires all pairs of traces to be compared to each other. Even using cascaded map-reduce jobs to reduce and pre-process the amount of data, the last reducer is left with this ultimate task. We facilitate computation of our results despite this restriction using the algorithms and sampling described in section III-C & IV.

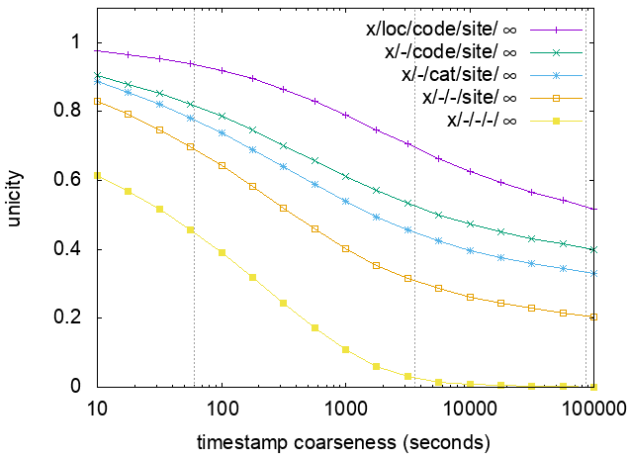

Fig. 2. Click trace unicity over coarsened time.

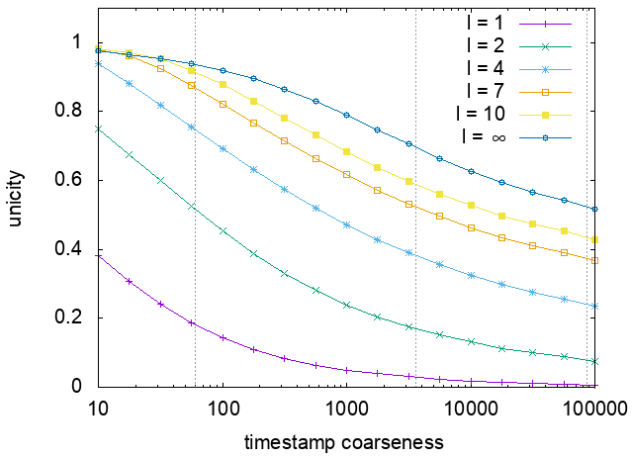

Fig. 3. Click trace unicity, trace length limited to l. Configuration: x/loc/code/site/·.

B. Applied Anonymization

Following the common argument of tracking companies, we anonymize the data by generalization (coarsening, truncating, omission) as described in Section III-D, and evaluate its effect on the perseverance of pseudonyms. We slightly need to adapt this step, given the dataset we have at hand.

We generalize with respect to four properties: the temporal resolution, client geolocation, information about the visited page, and finally the length of each click trace. Trace length can be adjusted by discarding traces below a minimum length or cutting traces above a maximum length into several, smaller ones. Page information at the highest level of detail consists of the site domain and a code. The code contains information about the exact page, which, for identification purposes, implies information about its category. We subsequently generalize to the tuple of site and category, thus generalizing the specific page to the category it belongs to. Then, we also omit the category and only consider the site, and finally we omit all information about the page.

Within experiments and results we denote which information is used by the tuple *<temporal resolution> / location / [code/category] / site / <trace length>* ("ms/loc/code/site/∞" for instance denotes information at the original granularity with click trace length corresponding to the sessions as defined above). Omissions are denoted by a dash "-": so "s/-/-/-/∞" represents a dataset with timestamps (coarsened to seconds) and trace length. The click traces may still be identifiable through timing and their length, but leaking such a trace could not disclose any information about the visited sites. At even lower granularity, "d/-/-/-/1" denotes a database containing only the information on what day each click occurred, without any information pertaining to client, site and sequence. If one of the elements is displayed on the x-axis of the plot, the field is correspondingly replaced with an "x". If the x-axis displays timestamp coarseness, we added dotted vertical indicator lines for coarseness values of a minute, hour and day (60, 3600 and 86400, respectively).

C. Unicity and Pseudonymity

Our first research question investigates to which extent pseudonyms emerge in tracking data, and how they are affected by successive generalization of the data.

1) General Unicity: We determine the unicity according to Algorithm 2 on the sample described in Section IV. It initially includes highly detailed attributes per click (location, code, and site, timestamps in ms), which we gradually coarsen as described above. While industry parameters for temporal coarsening are usually on the order of tens of minutes [4], we coarsen up to slightly more than a day to provide a more comprehensive overview. We first run a series of experiments on the entirety of the sample across all sites, and then repeat the series on the trace sets per site.

We expect high unicity in experiments with high levels of detail in auxiliary information and timestamps as well as no length restrictions. Following the common rationale that coarsening suffices to anonymize tracking data, we expect the unicity to decrease markedly in the subsequent experiments with lower granularity.

The results of the first series of experiments are shown in Fig. 2. We observe that unicity highly depends on timestamp coarseness. Reducing the accuracy of the timing to the order of seconds or minutes obscures the exact instant of a click, but details on intervals between page calls is retained in the data. When all information about client and page are removed, unicity remains at over 60% for high temporal resolutions.

In the cases where timestamps are coarsened to the order of hours, most differences in the intervals between clicks are lost, and just time of day and sequence information of the clicks are preserved. We can observe that the granularity of information about user and page still has a marked effect. At high granularity, taking information similar to the data contained in current tracking databases, over 70% of all traces remain unique. When removing the location and all page

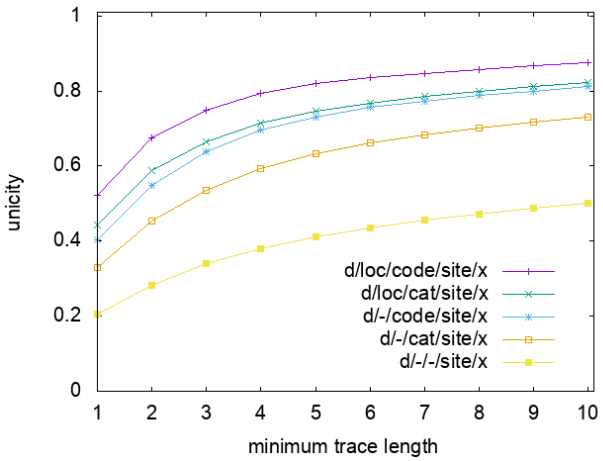

Fig. 4. Click trace unicity with minimum length, timestamps coarsened to the day.

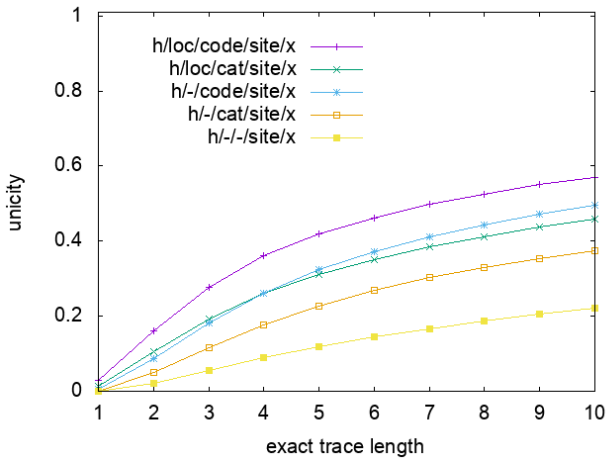

Fig. 5. Click trace unicity for exact trace length, timestamps coarsened to the hour.

information save the website domain, unicity is reduced to a value slightly below 40%.

Further coarsening timestamps to *the day of the click* removes differences in time zones and browsing habits pertaining to daily routines.With full information about page and location, over half of all click traces remain unique. Knowing only the number of clicks per day on the websites contained in a trace (day/-/-/site/∞), the unicity still remains at over 20%.

Only when deleting all auxiliary information (x/-/-/-∞), and coarsening the timestamps to the order of hours or days do we observe unicity values that converge to 0 – where practically no pseudonym has remained and the database is anonymized.

The results of this first series of experiments show that generalization of the attributes does not yield anonymity as long as even a minimal amount of utility of the database is preserved.

2) Trimming Sessions and Click Trace Length: Long click traces exhibit higher unicity than short traces and at session lengths > 50 clicks even strong coarsening has little effect. This is intuitive, as the attribute of a click trace's length is progressively characteristic with increasing length. Deliberately limiting click traces to a maximum length should therefore increase anonymity.

We perform a second series of experiments to this end, reducing the length of click traces but using all auxiliary information about clients and pages. Practically, this represents a tracker that forcibly resets sessions or client IDs after an observed number of clicks, to reduce the likelihood of pseudonyms to emerge.

The results are shown in Fig. 3. We observe that unicity can indeed be significantly reduced as the trace length l is reduced, particularly at high timestamp granularity. However, truly low values of unicity are only reached when traces are limited to a single or two clicks and timestamps are coarsened to at least several hours. Even keeping traces of just $l = 4$ clicks, and a time resolution of hours, is sufficient to uniquely identify over 40% of all click traces.

Furthermore, our dataset contains a number of session fragments - sessions where only one or two clicks occurred on a website in our sample - which arguably aren't representative for the types of browsing sessions we are concerned about. To assess their effect on the measured unicity, and to get a better insight into the unicity of click traces more likely to represent full browsing sessions, we performed another series of experiments with increasing minimum length. As short click traces are less likely to be unique, we expect unicity to rise when shorter click traces are removed. Therefore we only show the results for highly coarsened timestamps. First, we only remove shorter click traces and leave longer ones intact, coarsening to the order of days (Fig. 4). Next we cut click traces to a chosen length as in the previous experiment and subsequently remove all shorter click traces from the sample, coarsening to the order of hours (Fig. 5).

In both experiments unicity rises sharply by over 20% when click traces consisting of only one or two clicks are removed, confirming our expectation that unicity of full browsing sessions is likely higher than our initial results show.

3) Unicity of Local Tracking: We finally wanted to take the position of publishers that apply local tracking, meaning websites that keep log files containing the clicks of their own visitors. Our dataset contains all necessary data for all participating websites, as they share all calls to their pages with the ABC, to give an accurate picture of their popularity.

Click traces within single sites are bound to be much shorter than click traces of multi-domain sessions. The universe of different pages within a single site is also much lower, and a small number of specific pages or categories have been shown to be much more popular than others[17]. Given these points, we expect the unicity of click traces per site to be much lower, than of click traces from cross-domain tracking.

The results show that unicity does decrease, slightly. Within single pages, when the location of the client is removed and

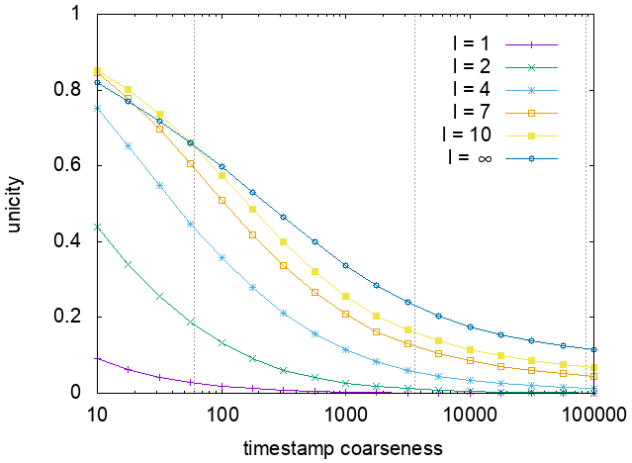

Fig. 6. The "Matomo (Piwik) case", local unicity at max length l. Configuration: x/-/cat/-/·.

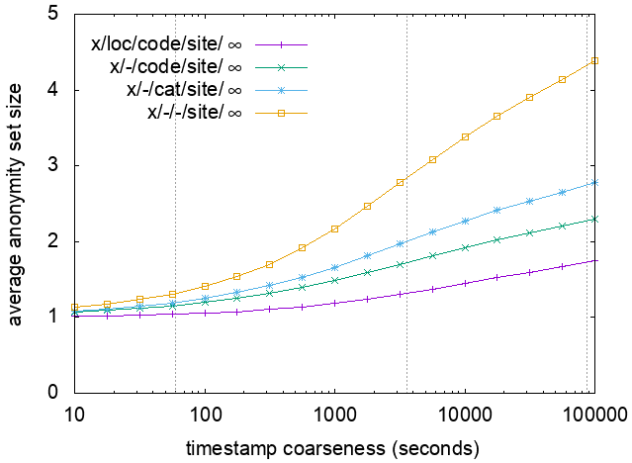

Fig. 7. Average anonymity set size within the original dataset.

only the category of the called page is retained, the unicity of longer click traces remains high (well above 20%), even when the temporal resolution is reduced to the order of hours (cmp. Fig. 6). Only limiting the trace length to a single click with timestamps on the order of minutes or hours, or to 2-click tuples with timestamps on the order of hours pushes unicity below 10%.

An interesting artifact can be observed in Fig. 6 for configurations of very high granularity (i.e. low timestamp coarseness, high maximum length): a decrease in maximum length can cause an increase in unicity. The reason for this lies in the way that sessions are extracted from the measurement databases: Using Algorithm 1, single longer sessions are split into several smaller click traces. For relatively high length limits ($l = 10$) and given information of high granularity, the odds are fairly good that a single, long unique click trace is cut into several shorter click traces, all of which remain unique. This in turn increases the overall unicity of the dataset.

4) Anonymity Set Sizes: Unicity measures the fraction of pseudonyms in the database, the remaining non-unique click traces fall into anonymity sets of varying size. While the number and cardinality of these sets has no effect on unicity, it has long been identified as a reliable metric for anonymity in settings such as DC networks [18] We therefore investigated the size of these anonymity sets using Algorithm 2.

For the sake of stable anonymization, one would strive for large anonymity set sizes. Given the large number of over a million clients and the common assumption of the popularity of pages being power-law distributed [19], many common, or at least highly similar behaviors should be contained in the database. Observing the impact of time granularity, one would expect to see fewer, larger anonymity sets, especially when reducing the temporal resolution, and hence the impact of different users starting their browsing sessions at different times. This should be pronounced for parameter sets that are already characterized by low unicity.

The results, as shown in Fig. 7, largely follow these expectations. Increased coarsening causes anonymity set sizes to continue to increase linearly even as unicity converges. This indicates indeed that with increased coarsening few, increasingly large sets of identical click traces emerge, whereas the fraction of unique behavior, probably clicks to rarely visited pages, or browsing with a highly identifying user agent characterization, remains comparatively stable.

D. Identifiability Experiments

Unicity provides a measure of the pseudonimity of click traces in a tracking database. It does not indicate how easily a click trace may be identified or how much an adversary could gain by doing so. If a click trace is only unique when considering the entire trace, then the pseudonym *is* all the data and knowing to whom it belongs does not offer any additional insight. Insight can only be gained when the trace contains unique subtraces.

Our second research question therefore aims at understanding how easily tracking data can be linked against, and thus re-identified with data from secondary sources. We devise a series of experiments to measure the identifiability, using Algorithm 3 as defined in Section III-C. We are interested in both of the scenarios described: the case of trackers partaking in user data exchanges and comparison to data that is publicly available (for instance on Twitter).

1) User Data Exchanges: In the first scenario, we consider a tracker to have collected a dataset using his own tracking technology and to acquire a second dataset at a user data exchange, such that the acquired dataset has a partial overlap of tracked sites with their own data. Given various percentages of overlap we (1) want to understand, how large a fraction of the click traces in the second dataset can be re-identified and uniquely matched to click traces in the first dataset. We subsequently are interested in (2) the *gain* of the adversary. Section III-C defines it as the fraction of identified clicks (contained

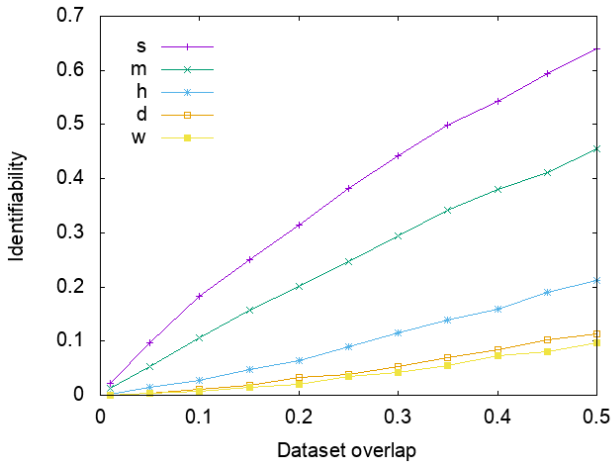

Fig. 8. Dataset enrichment via acquiring a dataset: We measure the identifiability of the acquired dataset given an overlap between the datasets. Configuration: ·/-/-/site/∞

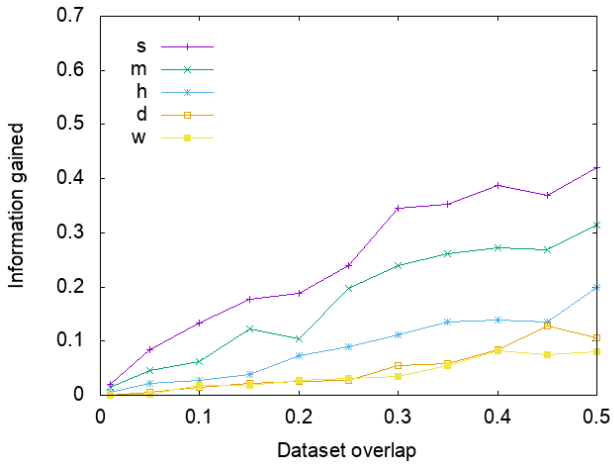

Fig. 9. Dataset enrichment via acquiring a dataset: We measure the information gain, meaning the amount of information in identified click traces not previously known, given the overlap between datasets. Configuration: ·/-/-/site/∞

in identified click traces) in an acquired dataset, which were not already contained in the initially owned dataset. This data can then be used to create more comprehensive click traces belonging to the same data subject even though the subject is not explicitly contained in either database.

We adapt Algorithm 3 to incorporate the prior knowledge of the adversary. For that purpose, we sample a *database overlap* by selecting a random collection of websites such that the number of clicks belonging to those websites is a given fraction of the overall number of clicks. We chose this approach due to the high variance in size of different websites in our dataset.

Geolocation, the public host part of a site, and on a somewhat reduced resolution even the times of clicks can be considered globally valid and compatible between tracking collections. Given that the acquired data is bought at a user data exchange, however, the adversary may not get, or not be able to interpret, the code nor category in the trace set, as these are schemes that are agreed upon between tracker and publisher internally. We thus remove these details from the datasets completely, before calculating the identifiability.

We expect identifiability to increase progressively with a growing overlap, eventually becoming equal to unicity once overlap reaches 100%. The results confirm this expectation (cmp. Fig. 8), showing an almost linear relationship between identifiability and overlap for overlap values below 0.5. However, as the overlap increases, the adversary's potential payoff decreases, because he can only learn new information in the non-overlapping portion.

We therefore turn to calculating the *gain* as defined in section III-C. On first glance, if unknown clicks are distributed uniformly across identified and unidentified traces, gain should be equal to identifiability. If, for instance, half of all click traces are identified, then we would expect roughly half of all unknown clicks to be contained therein. However, such an assumption of uniformity would be incorrect, because the chance of a click trace to be identified increases with the number of known clicks it contains. It is expected then, that the set of identified click traces contains a lower proportion of unknown clicks than the set of unidentified click traces. The results as shown in Fig. 9 confirm this expectation. At high granularity, an overlap of 30% enables the adversary to learn about 20%-35% of the unknown clicks and at very low granularity with coarseness values on the order of days, at 40%-50% overlap, about 10% of clicks remain susceptible to identification.

2) Shoulder Surfing and Comparison to Digital Dossier Aggregation: Finally, we want to assess how easily data subjects can be linked back against their click traces in tracking databases. For this purpose, we assume an adversary to possess some identified page calls of a user, known from secondary sources. Generally this may be from a direct physical encounter or prior knowledge of their browsing behavior, but observation of publicly posted links on social media or, more generally, digital dossier aggregation may be more attainable. In this case the adversary is usually aware of at least part of the data subject's identity and their goal is to use the observed information of a partial browsing session to discover the full browsing session in a database of web tracking data.

The categorization and mutual agreement on codes between tracker and publisher is assumed to be unknown to the adversary, we thus limit their best case knowledge to access time, location, and the visited website. Note that we do not assume that observations are consecutive. Each observation is selected completely at random from the browsing session, but the overall order is preserved. This corresponds to the definition of click traces, subtraces and identifiability as laid out in section III.

We perform the experiments on data of various coarsening

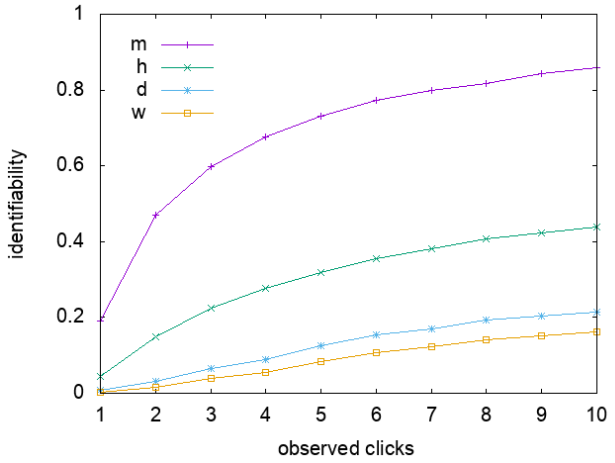

Fig. 10. Shoulder surfing: We measure the identifiability of a partially observed browsing session, given the number of observations. Configuration: ·/loc/-/site/10.

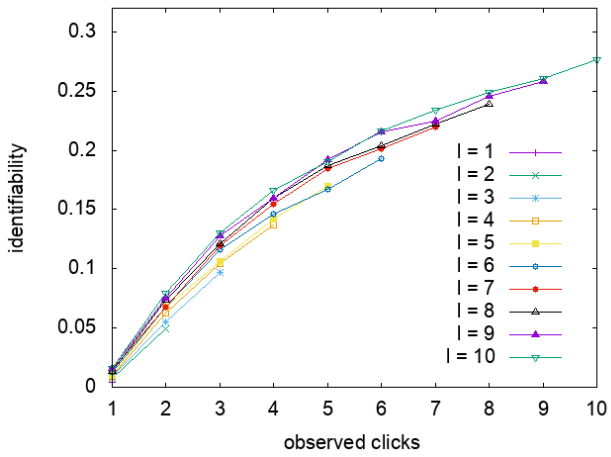

Fig. 11. Shoulder surfing: We measure the identifiability of a partially observed browsing session, given the number of observations for different session lengths. Configuration: h/loc/-/site/·.

levels and evaluate over the extent of prior knowledge of the adversary. Note that as the adversary gains more observations, click traces with an overall length below the number of observations are no longer considered. So on one hand we expect identifiability to be well below unicity, due to unicity acting as an upper bound for what can be identified. On the other hand, due to the removal of short, low unicity click traces, we would expect even a relatively short number of observations to be sufficient for a significant degree of identifiability.

The results in Fig. 10 indeed show that the adversary needs to know the time of a single visit only to an accuracy on the order of minutes to identify almost half of all browsing sessions with just two observations. Timestamp coarsening reduces identifiability substantially, but doesn't eliminate it. Knowing the hour of the observation as well as the website's domain and location of the user, four observations suffice to correctly identify a quarter of click traces. Once the adversary has made seven or more observations, knowing merely the day is sufficient to identify 20% of browsing sessions.

We finally want to assess whether trimming of click traces is as effective at reducing identifiability as it is at reducing unicity. Fig. 11 plots the identifiability of click traces trimmed to length l, given prior knowledge of a number of observations, while limiting the maximum length of click traces in the database. Longer click traces exhibit higher unicity, but not necessarily higher identifiability. This is because we assume the same adversary strength in both cases, translating to a relatively smaller observation size in the case of the longer click trace. Interestingly, these effects (increased unicity, but smaller subtrace) appear to cancel each other out for the most part. At the lower end of maximum length we observe a statistically significant but very small increased identifiability for higher length limits and at the mid and high end the results become almost indistinguishable.

VI. Related Work

We investigate the emergence of pseudonymous data and identifiability of click traces in databases of Web trackers. The primary research areas related to our work pertain to identifiability of online tracking and browsing data and sequential identifiability in general.

A. Online Tracking

Third party online tracking mechanisms, such as the ones which were used to generate our data, are widely used [3]. *Cookies* remain the most common form [1], and techniques such as "evercookies" and "cookie syncing" continue to make them a very resilient and reliable tracking tool [20]. *Browser fingerprints* are a more recent technique which attempts to identify users via information extracted from their browser. The potential of this approach has been demonstrated by Eckersley and others [21], [22], [23] by leveraging a combination of seemingly benign information to generate highly identifiable fingerprints. Besides such browser specific information, Upathilake et al. [24] identify additional categories of fingerprints: Based on Canvas [25], JavaScript Engine [26], and Cross-browser.

While our work is based on data obtained through online tracking, the data was gathered entirely through third party cookies and we do not generate fingerprints in the conventional sense. In fact we purposefully disregard a large amount of information that would traditionally be used to generate fingerprints to demonstrate that identifiability can be achieved in other ways.

B. Browsing History

Broadly speaking we analyze the metadata of browsing sessions. We thus implicitly analyze browsing behavior. However, research corresponding to browsing behavior usually analyzes how users tend to navigate through websites [27], rather than privacy in web tracking.

Olejnik et al. [28], [29] do examine the uniqueness of browsing history patterns and their work strongly relates to ours. However, their analysis of unicity diverges significantly from our work. They do not investigate unicity as it emerges in cross site tracking, but rather explore differences in repetitive browsing behavior, completely observing the users in all their actions. The effect of anonymizing, or coarsening measures is also not considered in their work.

Browsing history is also used by Su et. al [30] to link a data subject to their social media profile by scraping profile activity and matching the visited sites to the history. Again, this represents an attempt to identify profiles from a local observer's view, unlike our analysis of global tracking data.

Finally, Yu et. al [2] classify certain pieces of information as particularly identifiable (or "unsafe"). Their scale of evaluation is similar to ours, with a comprehensive overview of the German online sphere. However, while we focus our analysis on the anonymizability of the datasets, they took a more local approach and implemented a client based browser extension as a means of privacy protection.

C. Trace Unicity

The notion that sequences of data fragments leak private information has been explored before. Papadimitriou et al. [31] studied time-series compressibility and privacy using perturbation. Le Blond et al. [32] show that over 98% of VoIP calls in their dataset can be traced using call start and end times with 1-second granularity. Fan et al. [33] investigated if differential privacy techniques on sequential information protect sensitive data while retaining specific utility.

Re-identifying claimed, or seemingly anonymous data represents its own entire research. Naranayan and Shmatikov were able to successfully de-anonymize parts of an industry dataset, in their groundbreaking paper from 2006 [10]. This has spawned large interest, and several approaches to improve anonymization and re-identification have been published [34], [35], [36], [11]. This terminally led to the acceptance that only differential privacy can provide guarantees. We were inspired by this work. But given the industry practice of IP address truncation, and claims that such coarsening sufficiently anonymized their datasets, we were interested in the limits of this approach.

Similar to De Montjoye et al. [15], we approach sequence privacy through the upper bound of unicity instead. In their initial work they examine the unicity of location sequences to explore re-identifiability through movement patterns. They repeated a similar approach using credit card shopping data in [37]. When considering the unicity of sequential information, the generating process for that information has to be considered. So even though we look at browsing sessions in a similar fashion to De Montjoye's approach, the underlying model which generated the data is different and conclusions from one do not apply to the other.

VII. Conclusion

In this paper we have shown that sequential browsing data is highly identifiable and attempts to lower the identifiability through coarsening are largely ineffective.

We analyzed a dataset of browsing sessions representative for both local analytics on single websites as well as large cross domain trackers. We wanted to understand (1) how common it is for pseudonymous data to emerge in such databases, as privacy regulations require informed consent if pseudonyms are processed. And (2) how vulnerable such databases are to re-identification with partial knowledge in practical applications. Throughout this endeavor we wanted to know to which extent coarsening or generalization, the industry standard for anonymizing such data, helps to protect the privacy of the tracked audience.

Our results show that unicity, the prevalence of pseudonyms in the data, is very high for almost all configurations. Pushing unicity below a level of 10% requires removal of all information pertaining to clients and website visits, and coarsening timestamp information to at least an order of hours. We make no judgment which level of unicity is or is not legally acceptable. However, it stands to reason that unicity is highly indicative of how vulnerable such data is to re-identification, especially considering future capabilities, both regarding the processing of data as well as the amount. In the absence of more effective anonymization methods it appears very unlikely that any meaningful degree of utility can be preserved in a database of clicks without pseudonymous data.

In our practical evaluation of identifiability this idea is largely confirmed. Trackers that participate in the common markets of user data exchanges have to assume that large parts of the data they are passing on can immediately be re-identified by the recipients. Shoulder surfing attacks, or the knowledge of two to three visited Web pages, for instance from somebody's Twitter feed, are sufficient to uniquely identify entire browsing sessions retroactively. These results are consistent with and strengthen established research. Website visits where users considered themselves unobserved can easily be attributed to them as long as part of that visit was observed, even if the observer is restricted to the website's domain and rough location of the user.

Our results strongly imply that audience measurement providers who want to anonymize click traces in compliance with regulations such as the GDPR will need to use methodology beyond coarsening. Adding noise or otherwise perturbing the data, for example to achieve differential privacy, provides provable privacy guarantees at the cost of significant losses in utility. These methods have been applied in specific circumstances, but have not been widely adopted by audience measurement providers.

In summary, we observe that sequential browsing data contains highly identifiable information. Anonymizing such data by generalizing its attributes has little effect; even if session recording length is severely restricted and click traces are trimmed to only two or three page calls. According to our research, if negligible identifiability is desired, only single page calls with a minimum of additional information about the browser and the visited page can be stored.

REFERENCES

[1] F. Roesner, T. Kohno, and D. Wetherall, "Detecting and defending against third-party tracking on the web," in *USENIX conference on Networked Systems Design and Implementation*, 2012.

[2] Z. Yu, S. Macbeth, K. Modi, and J. Pujol, "Tracking the trackers," in *Proceedings of the WebConf (WWW)*, 2016.

[3] S. Englehardt and A. Narayanan, "Online tracking: A 1-million-site measurement and analysis," in *ACM CCS*, 2016.

[4] S. Passmann, A. Lauber-Roensberg, and T. Strufe, "Privacy-preserving audience measurement in practice—opportunities and challenges," in *IEEE Communications and Network Security (CNS)*, 2017.

[5] J. R. Mayer and J. C. Mitchell, "Third-party web tracking: Policy and technology," in *IEEE Security and Privacy*, 2012.

[6] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," technical report, SRI International, Tech. Rep., 1998.

[7] Ú. Erlingsson, V. Pihur, and A. Korolova, "Rappor: Randomized aggregatable privacy-preserving ordinal response," in *ACM CCS*, 2014.

[8] D. L. Quoc, M. Beck, P. Bhatotia, R. Chen, C. Fetzer, and T. Strufe, "Privapprox: privacy-preserving stream analytics," in *USENIX Conference on Usenix Annual Technical Conference (ATC)*, 2017.

[9] A. Narayanan and V. Shmatikov, "How to break anonymity of the netflix prize dataset," *arXiv preprint cs/0610105*, 2006.

[10] ——, "Robust de-anonymization of large sparse datasets," in *Security and Privacy, 2008. SP 2008. IEEE Symposium on*. IEEE, 2008, pp. 111–125.

[11] ——, "De-anonymizing social networks," in *Security and Privacy, 2009 30th IEEE Symposium on*. IEEE, 2009, pp. 173–187.

[12] P. Papadopoulos, N. Kourtellis, P. R. Rodriguez, and N. Laoutaris, "If you are not paying for it, you are the product: How much do advertisers pay to reach you?" in *Internet Measurement Conference*, ser. IMC, 2017.

[13] S. Coffey, "Internet audience measurement," *Journal of Interactive Advertising*, vol. 1, no. 2, pp. 10–17, 2001.

[14] E. Sy, C. Burkert, H. Federrath, and M. Fischer, "Tracking Users Across the Web via TLS Session Resumption," in *ACSAC*, 2018.

[15] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific reports*, vol. 3, p. 1376, 2013.

[16] M. Arlitt, "Characterizing web user sessions," *ACM SIGMETRICS Performance Evaluation Review*, vol. 28, no. 2, pp. 50–63, 2000.

[17] T. Strufe, "Profile popularity in a business-oriented online social network," in *ACM Workshop on Social Network Systems*, 2010.

[18] A. Serjantov and G. Danezis, "Towards an information theoretic metric for anonymity," in *International Workshop on Privacy Enhancing Technologies*. Springer, 2002, pp. 41–53.

[19] L. A. Adamic and B. A. Huberman, "The web's hidden order," *Communications of the ACM*, vol. 44, no. 9, pp. 55–60, 2001.

[20] G. Acar, C. Eubank, S. Englehardt, M. Juarez, A. Narayanan, and C. Diaz, "The web never forgets: Persistent tracking mechanisms in the wild," in *ACM CCS*, 2014.

[27] R. E. Bucklin and C. Sismeiro, "A model of web site browsing behavior estimated on clickstream data," *Journal of marketing research*, vol. 40, no. 3, pp. 249–267, 2003.

[21] P. Eckersley, "How unique is your web browser?" in *International Symposium on Privacy Enhancing Technologies Symposium (PETS)*. Springer, 2010, pp. 1–18.

[22] A. Gómez-Boix, P. Laperdrix, and B. Baudry, "Hiding in the crowd: an analysis of the effectiveness of browser fingerprinting at large scale," in *WWW 2018: The 2018 Web Conference*, 2018.

[23] A. Vastel, P. Laperdrix, W. Rudametkin, and R. Rouvoy, "Fp-stalker: Tracking browser fingerprint evolutions," in *IEEE S&P 2018-39th IEEE Symposium on Security and Privacy*. IEEE, 2018, pp. 1–14.

[24] R. Upathilake, Y. Li, and A. Matrawy, "A classification of web browser fingerprinting techniques," in *New Technologies, Mobility and Security (NTMS)*, 2015.

[25] K. Mowery and H. Shacham, "Pixel perfect: Fingerprinting canvas in html5," *Proceedings of W2SP*, pp. 1–12, 2012.

[26] M. Mulazzani, P. Reschl, M. Huber, M. Leithner, S. Schrittwieser, and E. Weippl, "Fast and reliable browser identification with javascript engine fingerprinting," in *Web 2.0 Workshop on Security and Privacy (W2SP)*, vol. 5, 2013.

[28] L. Olejnik, C. Castelluccia, and A. Janc, "Why johnny can't browse in peace: On the uniqueness of web browsing history patterns," in *5th Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs 2012)*, 2012.

[29] ——, "On the uniqueness of web browsing history patterns," *Annals of telecommunications-annales des télécommunications*, vol. 69, no. 1-2, pp. 63–74, 2014.

[30] J. Su, A. Shukla, S. Goel, and A. Narayanan, "De-anonymizing web browsing data with social networks," in *The WebConf (WWW)*, 2017.

[31] S. Papadimitriou, F. Li, G. Kollios, and P. S. Yu, "Time series compressibility and privacy," in *International conference on Very large data bases (VLDB)*, 2007.

[32] S. Le Blond, D. Choffnes, W. Caldwell, P. Druschel, and N. Merritt, "Herd: A scalable, traffic analysis resistant anonymity network for voip systems," in *ACM SIGCOMM Computer Communication Review*, vol. 45, no. 4. ACM, 2015, pp. 639–652.

[33] L. Fan, L. Bonomi, L. Xiong, and V. Sunderam, "Monitoring web browsing behavior with differential privacy," in *The WebConf (WWW)*, 2014.

[34] L. Backstrom, C. Dwork, and J. Kleinberg, "Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography," in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 181–190.

[35] A. Gkoulalas-Divanis, G. Loukides, and J. Sun, "Publishing data from electronic health records while preserving privacy: A survey of algorithms," *Journal of biomedical informatics*, vol. 50, pp. 4–19, 2014.

[36] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proceedings of the National Academy of Sciences*, p. 201218772, 2013.

[37] Y.-A. De Montjoye, L. Radaelli, V. K. Singh *et al.*, "Unique in the shopping mall: On the reidentifiability of credit card metadata," *Science*, vol. 347, no. 6221, pp. 536–539, 2015.