Journal of Privacy and Confidentiality (2011) **3**, Number 1, 3–26

Analysis of Google Logs Retention Policies

Vincent Toubiana* and Helen Nissenbaum†

Abstract. To preserve search log data utility, Google groups search queries in log bundles by deleting the last octet of logged IP address. Because these bundles still contain identifying information, part of these search logs can be de-anonymized [27]. Without an external audit of these search logs, it is currently impossible to evaluate their robustness against de-anonymizing attacks. In this paper, we leverage log retention policy ambiguities to show that quasi-identifiers could be stored in sanitized search query logs and could help to de-anonymize user searches. This paper refers to Google Search and Google Suggest log retention policies and shows that even with the highest degree of anonymization that Google offers, one could separate user queries with a high granularity. Because Google Suggest is queried every time a user types a character in the Google Chrome navigation box, the privacy of Chrome users could be compromised with respect to their browsing histories. Such ambiguities within log retention policies are critical and should be addressed, as anonymized logs could be shared with third parties without prior user consent.

1 Introduction

Due to the large amount of data search engines collect, search log retention policies have been frequently criticized [1, 29]. Google search is certainly the web service that is subject to the most attention from privacy advocates, and such attention is justified by the influence that Google Search has on the market. Indeed, with more that 70% of market shares in July 2010 [23], Google continues to dominate the search industry. Being the leader in the search market, Google has spearheaded changes to search privacy settings that have occurred over the last few years among its competitors. For instance, Google was the first search engine to sanitize its search logs, and set an example for other search engines, like Microsoft and Yahoo!, which later sanitized their logs and provided eventually more guarantees [31].

In 2007, the Article 29 Data Protection Working Group (G29) asked major search engines to not store personal identifiers in search logs for more than six months [1]. Although most search engines did not comply with this request, all of them agreed to reduce the retention period of some personal identifiers. To preserve search query log utility, some search engines decided to just remove or alter certain bits of main identifiers. Google, for instance, modifies the last IP address byte after 9 months and deletes this last byte after 18 months [10]. This process is called generalization and, if used correctly, can guarantee reasonable privacy protection [36]. In some circumstances (only one IP in the bundle used to search), though, this generalization fails to provide

*Work done while a Postdoctoral Researcher at New York University. Alcatel-Lucent Bell Labs, Application Domain, mailto:vincent.toubiana@alcatel-lucent.com

†New York University, Media, Culture, and Communication, mailto:helen.nissenbaum@nyu.edu

any privacy guarantees. Additionally, Google anonymizes cookies that have been stored in the logs for more than 18 months. However, Google's privacy policy is very broad in the definition of this process [19].

While much attention has been paid to the log retention period [12], the effectiveness of Google log sanitization process has not be investigated so far, partly because the cookie sanitization process that Google applies to its search logs has never been officialy detailed. Google often defines its sanitization process by using verbs such as 'anonymize' [21, 6, 13, 13, 14] and 'obfuscate' [7]. However, these verbs describe the process's expected effects, rather than the process itself, thus making Google the only entity able to judge its own sanitizing processes. The ambiguity of the definition of this process may lead to different interpretations. Especially, *anonymization* [6] and *obfuscation* [7] seem to be defined by Google as the modification or deletion of bits of some identifiers [6, 34] and may not totally remove quasi-identifiers. As a result, in May 2010, the Article 29 Data Protection Working Group asked for an external audit of Google's anonymized search logs [29].

In this paper, we suggest that quasi-identifiers [36] could exist in the sanitized search logs. These quasi-identifiers are composed of portions of the cookies that are not deleted or modified, User Agent strings and generalized IP addresses. In their analysis, [27] used only the generalized IP address in conjunction with Machine Learning techniques to identify fibers in the anonymized logs. However, if the cookie anonymization process does not mask quasi-identifiers in the *Cookie ID*, it could be very simple for an attacker to de-anonymize the search logs. An attacker could use these quasi-identifiers linking all the searches made by a user and then connect them to an arbitrary pseudonym. This process, that we call **pseudonymization**, is a first step to de-anonymize search logs. Indeed, once the logs are pseudonymized an attacker just has to extract personal identifiers [26, 27] in these search logs to replace pseudonyms with personal identifiers.

To be effective a log sanitization process should mask every quasi-identifier to break the links between user searches, and eventually provide a guarantee similar to k-anonymity. However, due to a lack of documentation about the cookie anonymization process, it is not possible to evaluate its real impact.

Therefore, we investigate the potential use of the sanitized *Cookie ID* to define a quasi-identifier in Google sanitized search logs. Based on the documents that Google provides about the cookie anonymization process, we show that the *Cookie ID* could have remained unique after sanitization. Indeed, in all the official documents we found, Google PREF cookie is obfuscated (see [7]); anonymized [21, 6, 13, 13, 14]) or modified (see [34]); its unique ID number is deleted (see [6]) but the cookie deletion is never mentioned. Although Google has never disclosed sanitized search logs publicly, its privacy policy allows the company to share search logs with third parties without first requesting user consent [19].

The rest of this paper is organized as follows. Section 2 quotes relevant sections from Google Search and Google Suggest privacy policies. This section also refers to additional documents explaining Google's log sanitization process. In Section 3, we reconstruct a picture of what might be a typical search log entry before and after it is sanitized.

Section 4 describes quasi-identifiers that remain in the sanitized logs. In Section 5, we use the disclosed AOL search logs to show that most searches in Google log bundles may be pseudonymizable. Section 6 discusses access to these 'anonymized' logs with regard to Google's privacy and internal policies. Section 7 overviews the existing log sanitization algorithims. Finally, Section 8 highlights the ambiguity of these sanitization processes and offers recommendations.

2 Google Retention Policies

Google's Search logs sanitization process has changed twice during the last three years [13, 14]. However, each of these changes has been announced without providing any explanation of the sanitization process itself. These announcements mostly concern the log retention period and mention identifiers that are modified. In this section, we summarize these announcements and quote documents that mention search log sanitization. Also in this section, we discuss retention policies of Google suggest that are equally ambiguous.

2.1 Google Search Logs

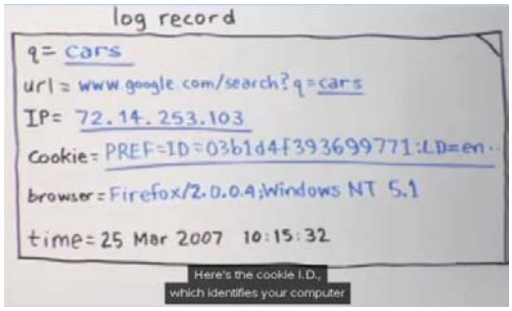
By searching the web for an explicit definition of Google's search log sanitization process, we found five kinds of documents that refer to this process:

1. Google Policy Update
2. Google Privacy Channel videos
3. Public responses to inquiries
4. Official slideshow used by Google
5. Scientific publications

Google Policy Update

In March 2007, following U.S and E.U. recommendations, Google was the first search engine to announce that it will anonymize its search logs [14]. Although the G29 asked search engines not to keep Personal Identifiers in their logs for more than 6 months [1], Google announced that its search logs would be anonymized within 18 to 24 months. In April 2008, Google decided to anonymize its logs after only 18 months [13].

The latest update in its search log retention policy has been announced in September 2008. Google made the announcement in its official blog, which claimed that it would anonymize "addresses on [our] server logs after 9 months". Google provides additional information on its privacy center website:



(a) Search log entry

(b) 18-month anonymized search log entry

Figure 1: Snapshots from Google's Privacy Channel video

> "We believe anonymizing IP addresses after 9 months and cookies in our search engine logs after 18 months strikes the right balance [...]"

Google Privacy Channel videos

In order to clearly explain its privacy policy, Google published a video explaining, in simple terms, the Google Search policy [21]. This video—the first of a series explaining how Google privacy policy applies to different services—was made in 2007, when search logs were sanitized after 18 to 24 months. Figure 1(a) is taken from the video describing the information Google collects when a user makes a search.

The video explains that the *Cookie ID* "identifies [a user's] computer and tells Google a [user's] preferences". A Google PREF cookie is detailed in its entirety in Figure 2. The video clearly shows that the *Cookie ID* is the entire string of character starting with 'PREF=ID', that includes the default language (LD=en) and the values that are masked by the ellipses: the number of results (NR), the two timestamps (LM and TM)

Name	Example	Description
ID	03b1d4f39369971	Cookie ID number
LD	en	Default Language
NR	10	Number of results
TM	1271285349	Timestamp of the cookie creation
LM	1271689668	Timestamp of the last preference changes
S	pf4EyV4GvtJp_Wx2	This value has never been documented by Google. It seems to be a hash of the precedent values

Figure 2: PREF Cookie full description

and the value of S. Although the values *TM*, *LM* and *S* are not defined to be unique, they may uniquely identify a browser.

Because *TM*, *LM*, *S* and the other preferences are not defined to uniquely identify a browser, Google also assigns a unique ID number that is stored in the PREF cookie. The cookie numbers are generated to assure a cookie's uniqueness. In the log record illustrated in Figure 1(a), the *Cookie ID* number is '03b1d4f393699771'. Although this *Cookie ID number* is only a small part of the *Cookie ID*, the anonymization process seems to focus only on this part of the cookie and may not alter other values. The distinction between the *Cookie ID* and the *Cookie ID number*, while subtle, is critical to understanding the cookie anonymization process described [6].

Figure 1(b) is taken from the same video, in which the anonymization process is mentioned. This snapshot shows how IP addresses and cookies are modified in the logs. The sanitized log entry that is illustrated in Figure 1(b) does not show whether or not the entire *Cookie ID* is deleted or modified during sanitization. Indeed, the ellipses that mask these fields are not replaced by 'X' but are simply not represented in the second snapshot.

Public responses to inquiries

After Google announced its log retention policy update in 2008, C. Soghoian asked Google for details about the log sanitization process. He then published the following response from Google [34]:

> "After nine months, we will change some of the bits in the IP address in the logs; after 18 months we remove the last eight bits in the IP address and **change the cookie information** (emphasis our own). It is difficult to guarantee complete anonymization, but we believe these changes will make it very unlikely users could be identified."

Soghoian concluded that if users do not delete their cookies, the 9-month sanitization process would not provide them privacy guarantees. Furthermore, this response does not

explain cookie sanitization and suggests that cookies are modified, rather than deleted.

The cookie alteration process is not detailed in the video [21] nor in Google's privacy policy. To the best of our knowledge, the cookie modification process was first mentioned in a letter sent to a congressman [6]. In this letter Google first describes the collected data as follows:

> "We collect typical log data, such as [...] the IP address [...]. We also may collect a unique *Cookie ID* generated for the computer from which the query originated. The cookie is used to recognize the user preferences... Google will anonymize the *Cookie ID* and the last octet [...] of the IP address associated with search queries after 18 months"

Google gives a description of the cookie anonymization process:

> "We currently plan to begin anonymizing PREF cookies [...] by deleting each PREF cookie's unique ID number."

The difference is subtle, but the *Cookie ID number* is deleted whereas *Cookie ID* (or PREF Cookie) is anonymized. In this letter, Google details the process in the following terms:

> "We plan to anonymize cookies that form part of our server search logs by deleting the ID number of each such cookie. For example, if a cookie that we serve to a computer at the time that a user searches with Google Search from that computer is assigned the number 740674ce2123e969, then 18 months after the search query is entered we plan to delete that unique Cookie ID number in its entirety."

Therefore, most information stored in the cookie may not be modified when search logs are sanitized; only the *Cookie ID number* is deleted. Although the remaining information may not uniquely indentify a user, it could certainly identify a user in a bundle. This is partly acknowledged by Google in its log retention policy FAQ [16]:

> "Cookie anonymization makes it less likely that a cookie can be used to identify a user."

Because *Cookie ID* is never deleted, it remains possible to link user searches based on other values that a cookie contains.

In a more recent document, P. Fleischer describes Google sanitization process with more details [11]. The implemented process seems to be slightly different that what Google planed previously.

> "Google decided to delete the last two digits from the IP addresses and alter the cookie numbers in our logs permanently after 18 months. This

> breaks the link between the search query and the computer it was entered from.[...] Here is what an IP address will look like in our logs after 18 months: 123.45.67.XX. After the same time period, the cookie will be replaced by a newly-generated cookie number."

This process clearly aims not to break search linkability but, instead, prevent linkage between a search made 18 months ago to searches that the user is currently making. There is no information about this newly generated cookie information, and it might be partially linked to the anonymized cookie and may even refer to some values that the original cookie contains. The paper later provides additional details about the cookie value:

> "At that point we permanently delete the last two digits from the IP address and randomly assign a new cookie number."

It seems that the new assigned number is randomly chosen, though the remaining information in the cookie may not be replaced by random values. Furthermore, the same number can be used to replace the same user cookie several times. A Google employee interviewed by D. Sullivan [35] provided more technical details:

> "we anonymize *Cookie ID* entifiers by transforming the identifiers with an HMAC (keyed hash function) using a randomly generated, ephemeral key for each day of logs, destroyed immediately after anonymization. This assures that it is impossible for Google to associate the cookie when it appears again with the old anonymized cookie, and it further assures that two anonymized cookies originating from the same cookie in two different days cannot be matched by Google."

Nevertheless, the fact that the anonymized cookie can link the searches made by a user on a day is confirmed in two sources [8, 15]. Therefore, this process is very likely to be the one that Google implemented to sanitize its search logs. Furthermore, this interview was published after Google developed the process, while the previous declarations were made when Google was still working on the technical details.

In this declaration, it was said that the entire *Cookie ID* is replaced. Consequently, the timestamps (*TM* and *LM*) are probably no longer stored in the anonymzed cookies. However, if a user is frequently doing a search that uniquely identifies her in the search log bundle, the cookie could be used to link all the searches made on one day, and this specific search could be used to link search history on different days. Since the PREF cookie is sent to every Google service, any request made every day to Google could be used as a way to link searches on different days.

Official slideshow used by Google

In an official slide show describing Google's Approach to Privacy [4], Google details a log entry corresponding to a search for the term 'flowers' (Figure 3(a)). This log

(a) Search log entry

(b) Truncated cookie description

Figure 3: Google's approach to Privacy slides 5 and 6

entry suggests that Google records the search URL, the major browser version, and a *Cookie ID* that is only represented by an ID number. A similar example showing an entry corresponding to a search for 'car' is used in Google's Privacy Policy FAQ [19]. These examples suggest that only a few pieces of information are collected when a user searches using Google.

In fact, this entry could correspond to a search made by a browser that did not have a PREF Cookie. In such a case, the server cannot record the cookie but may still record the freshly assigned *Cookie ID number*. The FAQ example comes with a notice clearly explaining that a cookie can be rejected or deleted. Furthermore, this presentation seems to illustrate log entries, rather than describe them exhaustively. For instance, the search URL does not include any parameter except the searched keywords. A search URL could contain many more parameters, as the log entry reconstituted in Figure 4 illustrates. Furthermore, in slide 6 (Figure 3(b)) the Cookie information is obviously truncated. Indeed, the S value is not shown in this cookie, and the displayed last character of the cookie string is a colon, while the last character of a cookie string should be a semicolon. This omission confirms that this presentation does not aim to be exhaustive. Notice, finally, that while the ID number is presented as a '*Cookie ID*' in slide 5, it is presented as the 'Randomly assigned ID number' in slide 6, thus suggesting that the two values can be different.

Scientific publications

In a recent technical report [8], a Google employee explains that anonymized server logs are used to evaluate the propagation speed of browser updates. The authors of the paper

show that the full browser version is recorded when a user makes a search on Google. They also confirm that the entire PREF cookie is recorded for every search, as they eliminate duplicate visits on the same day by counting "only the first Web request by each Web browser with the same Google PREF cookie". A previously published paper [15], whose authors confirm that cookies are much more accurate than IP addresses to identify browsers, and that only a small fraction of users reject or delete cookies, shows this. This fact also confirms that some Google employees have access to the anonymized logs. We discuss anonymized logs access policies in more detail in Section 6.

2.2 Google Suggest logs

Google Suggest is a service provided by Google to help Google Search users. This service, enabled by default on the Google page, analyzes the query that a user is typing and suggests related terms. Casteliuccia et al [3] showed how this personalized feature could be misused to recover part of a user's search history. Although Google has addressed this flaw, some privacy concerns remain since this service sends a significant volume of information to Google servers.

Video of Google Chrome by Microsoft

Every major web browser supports search suggestions, but the integration of this service in Google Chrome is more problematic. Unlike other browsers, Chrome does not separate the search box and the navigation box. Instead, it comes with a unique text box called the OmniBox that is used to both type a URL and to search.

While such an approach might be convenient, users can no longer type the URL of a website or search a page in the browser history without sending suggestion requests to the default search engine. Indeed, with Google Chrome, almost every keystroke in the OmniBox triggers a request to a search suggestion service (Google, by default) unless the service is turned off.

In a video that introduces Internet Explorer 8 functionalities [33], LePage emphasizes this privacy issue. LePage has been criticized for not mentioning the Google Suggest log retention policy [33].

Google Suggest log retention policy update

A few days after the release of the first version of Google Chrome, Google Suggest's retention policy was updated [25]. Some Chrome users expressed their concerns when they discovered that the browser was sending requests to Google servers. To address these concerns, Google updated the Suggest log retention policy:

> "For 98% of these requests, [Google] [doesn't] log any data at all and simply return the suggestions. For the remaining 2% of cases (which [Google] select[s] randomly), [Google] [does] log data, like IP addresses, in order to

> monitor and improve the service. [Google] anonymize[s] [this information] within about 24 hours in the 2% of Google Suggest requests [they] use."

Although Google claims that they anonymize 2% of the information requests they log, they do not provide detail to explain how these logs are anonymized. After Google announced the Google Suggest log retention policy update, privacy advocates have expressed concerns [28] due to the lack of clarity in the definition of this process.

Google Chrome privacy whitepaper

Google's suggest sanitization process has only been detailed very recently, more than a year after the retention policy update [18]:

> "IP address and certain cookies are also typically sent to your default search engine with the request. If Google is set as your default search engine, rather than logging all requests, Google only logs a randomly selected 2% sample of requests in order to help improve the suggestion feature and to prevent abuse. To preserve privacy, Google anonymizes the requests in these logs by dropping cookies and the last octet of the IP address within at most 24 hours."

It is worth noting that Google does not claim to drop all the cookies. Therefore, they could keep some cookies in the anonymized logs. In fact, sanitized suggestion logs may be similar to the sanitized search logs, meaning that these logs could contain the three first bytes of the IP address, the user agent, and a portion of the PREF cookie (the dropped cookies could refer to other cookies sent with the request).

Although the recorded logs concern only 2% of suggestion requests, these logs certainly contain a tremendous number of queries. Indeed, a suggestion request is sent after almost every keystroke in the OmniBox—including when the user is not searching. Even if a user just wants to browse cached web pages, or retrieve the URL of a previously visited website, Chrome sends requests to Google Suggest. Because Google Suggest queries are frequent (one for almost every character entered in the box), even 2% of them can still contain valuable information about user browsing behaviors. Finally, unlike browsing activities that are tracked by Google when the user is visiting a site belonging to the Google Content Network, user-typing activities are now reported to Google even if the third party cookies are blocked.

Google Instant retention policy

Google released 'Google Instant' in September 2010 to help users search faster by displaying results returned to search suggestions. This feature is already replacing suggest on Google Search and is likely to also replace it in Chrome. While the 'Google Instant' log retention period (two weeks [12]) is higher than the 'Google Suggest' log retention

period (one day), all Google Instant partial query data are deleted within two weeks [25].

3 Interpreting Google documents

Although we searched for a clear definition of the sanitized logs on all the documents that are available on the web, we were not able to find it. Nevertheless, from the documents that we retrieved, it is possible to draw a picture of Google sanitized logs. Notice that we affirm not that Google sanitized logs are exactly as we depict but that Google could keep such logs and still be in total compliance with its privacy policy and previous statements.

First, we describe what a record in Google Search logs looks like before being sanitized (see Figure 4(a)). This entry contains only un-authenticated data, meaning that such an entry is created even if the user is not logged into a Google Account. Information extracted from Google Search logs by [15] confirms that Google records the complete User Agent string; this is why this string is reported in its entirety.

Recently, the *Cookie ID number* size increased from 8 to 16 bytes. Now, this number is composed of two eight-byte values: *ID* and *U*. Another undocumented cookie value is GM. However, this value always seems to be set to '1' and, therefore does not carry any bit of identifying information. There are several other options that are not documented and that sometimes appeared in PREF cookies, but without access to search logs, we cannot estimate their uniqueness.

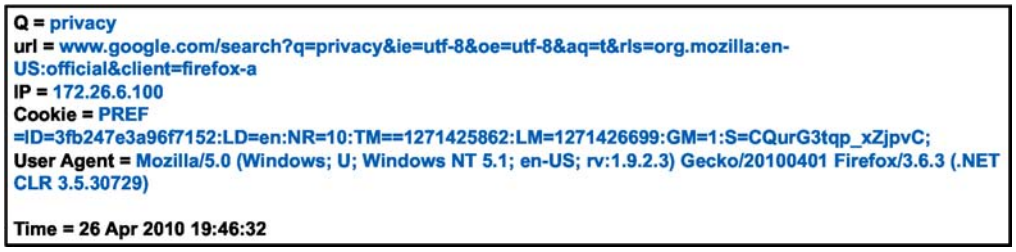
In Figure 4(b), we present a picture of a search log entry sanitized following the process described in [6]. The *Cookie ID numbers* are removed, and the other information is not modified. According to official description of Google sanitization process, the entry illustrated in Figure 4(b) could be stored for an indefinite period of time in search logs.

4 Retrieving Information in Anonymized Logs

In this section, we detail the attributes that compose the quasi-identifiers that remain in the sanitized search logs and explain how under certain circumstances, they can be used to re-identified.

4.1 IP Address

Even after 18 months, search and suggest query logs still contain quasi-identifiers that could be used to de-anonymize user queries. With the last byte of the IP address removed, the IP that was used to issue the request could be any of the 256 IP addresses in the block. However, for this process to be effective, a significant part of the users in the same address block should use the same search engine. Indeed, if a search engine counts only one active user in an IP address block, removing the three first bytes of the



(a) Search log entry

(b) 18-month anonymized search log entry

Figure 4: Reconstituted log entries

address will not help in preventing his or her re-identification. The impact of the IP modification is more effective if the search engine is very popular in the area covered by the IP block. For instance, in 2007, a block of 256 IP addresses contained on average 57 Yahoo! active users [27].

4.2 User Agent

According to a recent experiment conducted by the EFF [9], a browser version carries, on average, 10.5 bits of identifying information. A consequence of the experiments results is that, on average, only two browsers in a group of $2^{10.5}$ ($\approx$ 1,500) will send the same User Agent string. This result does not mean that in any group of 1,500 browsers, a browser version is unique; some browser versions are more popular and some are very rare. Users of popular browser versions bear less risk of being re-identified than do those of more rare versions. In a bundle grouping the searches issued from 256 distinct IP address, a given browser version may be unique and could eventually be a quasi-identifier.

Although the User Agent string is not fixed over time, the Operating System (OS) and browser may not be updated simultaneously, thus a link may still exist between successive anonymized searches log entries. Even if a user changed its operating system and browser simultaneously, her searches could still be linked if she is the only person in the bundle to have updated both the system and browser.

Because the User Agent string is browser-specific, it does not help to separate the searches that have been made by different users on a same computer. Furthermore, it is not unlikely for the browser version to carry only a few bits of identifying information once queries are grouped in bundles. Indeed, browser and operating system popularities are subject to geographical preferences. Specifically, employees of a large corporation may be constrained to use a specific browser version. In this particular situation, the user agent will not help to pseudonymize a bundle of queries.

4.3 Google PREF Cookie

Google search query logs may also contain information provided by the unaltered part of the PREF cookies. Indeed, no official Google documents mention the deletion of this cookie in its entirety. The cookie is obfuscated (see [7]; anonymized [21, 6, 13, 13, 14]) or modified (see [34]); its unique ID number is deleted (see [6]) but the cookie deletion is never mentioned. Most values in the PREF cookie seem fixed and do not change unless the cookie is removed. If it was stored in the sanitized cookie, the combination of *TM* and *LM*, could be used to pseudonymize search logs. A sanitized cookie may still carry other pieces of information like the user language and the number of results that should be displayed on a page.

5 Pseudonymizing Search Query Logs

We refer to the process of 'pseudonymization' as the process linking all the searches made by a user and then connecting them to an arbitrary pseudonym.

In this section, we explain what can be done if Google just applies the sanitization process that is mentioned in its official declaration. The timestamps are very revealing but there is evidence from an anonymous source [35] that the value of *TM* and *LM* are at least truncated if not removed from the sanitized cookies. However, the process currently described in official documents may not be effective.

This section shows that a large part of the sanitized search logs, corresponding to users that do not delete their cookies, could have been pseudonymized if Google just applied the process described in official documents. Then, we investigate how complementary pieces of information available in query bundles could have been used to pseudonymize the searches of users who frequently delete their cookies.

5.1 Long term cookie distribution

To evaluate the uniqueness of the value of *TM* and *LM*, we performed an experiment: we retrieved more than a thousand Google PREF cookies from the web, and then observed the uniqueness of some cookie values.

Obtaining PREF Cookies

These cookies have been obtained by querying Google with a string of characters containing the PREF cookie key names. Because Google limits the number of returned documents to 1000, and because some cookies appeared on multiple websites, only 1,069 unique cookies have been retrieved. The 1,000 most recent cookies have been kept for this evaluation.

Some retrieved cookies belonged to users who published them on forums and blogs, some may have been issued for third party servers that publish content processed by Google services. In fact, we noticed that some of these cookies appeared in websites that publish content translated by Google Translate. This content is published directly on these websites, and the result of the request might not be parsed appropriately. Consequently, these websites display the HTTP headers (including the cookie) instead of the translated content.

According to the timestamps, the oldest cookie was issued in May 2005, and the most recent was issued in May 2010. Figure 5 illustrates the time distribution of these cookies. Notice that more than a half of these cookies were issued within the last six months.

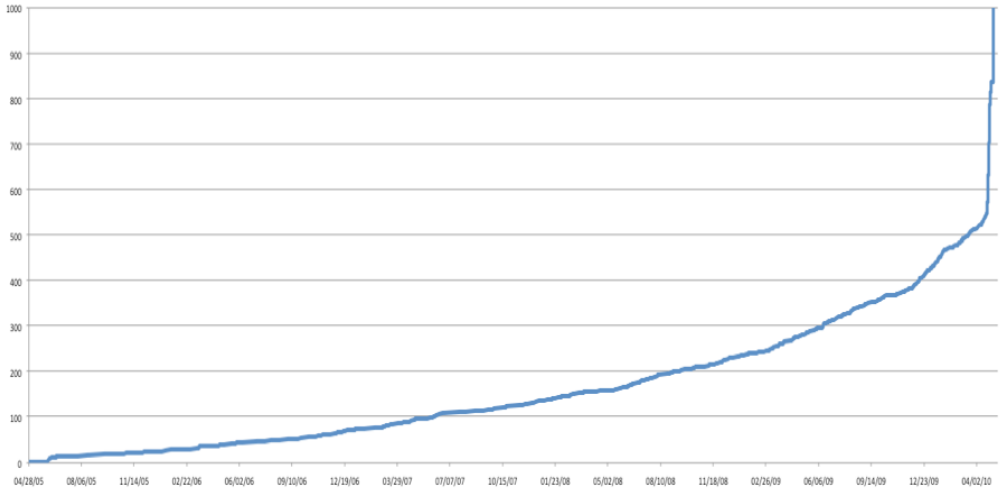


Figure 5: Web retrieved cookies timestamp distributions

Cookie uniqueness

We searched for timestamp collisions in this set of 1,000 cookies and detected four collisions of the cookie creation timestamp *TM*. Each collision involved only two different cookies. Interestingly, in one of these collisions, the cookies were differentiable based on the cookie modification timestamp *LM*. For one of the cookies, *LM* was one second behind *TM*. This difference is certainly caused by a slight latency between the two value computations. Although our sample might be too small to accurately estimate the timestamp collision probability, these results show that the combination *TM*, *LM* can be used as a quasi-identifier linking all of the searches made by a single user.

Cookie modifications

The number of cookie modifications was also observed and only 194 cookies had different values of *TM* and *LM*. For 123 of these cookies, the difference between *TM* and *LM* was less than 10 seconds and was certainly caused by latency. Therefore, on the 1,000 retrieved PREF cookies, only 71 reflect user preference changes. This low value of preference changes is explained partly by the method used to retrieve these cookies and the fact that some cookies actually belong to the server using Google services.

This phenomenon can also be explained by the localization that Google provides based on user IP addresses. With this localization by default, the preference of the *LD* is not frequently used. Furthermore, users can specify preferences for a search session either by adding parameter directly in the search URL or through the 'advanced search' link.

Recall that the hash value S is updated only if the value of one of the cookies changes. Because these values do not change frequently, the value of S that corresponds to a cookie is very unlikely to change. Since this value takes into account the ID number, it provides a direct match that recognizes most searches that a user has issued. If the user modifies a cookie, he or she can be reconnected to previous searches using the timestamp (TM, LM). Therefore, the hash value S is certainly considered as an identifier and removed (or replaced) in sanitized search logs.

This result also shows that more than 99% of the PREF cookie does not carry any preference information but has only an identifying purpose. If the purpose of the PREF cookie is mainly to store user search preferences, this cookie could be set only when a user modifies the default search preferences.

5.2 Large Network Cookie Distribution

In this section, we evaluate the uniqueness of a cookie timestamp based on the AOL search logs. For every user whose searches have been disclosed, we keep only the timestamp of the first search in these logs (as it corresponds to the definition of the TM value stored in Google cookies) and then evaluate the collision probability. Because these logs have been recorded over a relatively short period of time, the collision probability should be higher than for regular search logs.

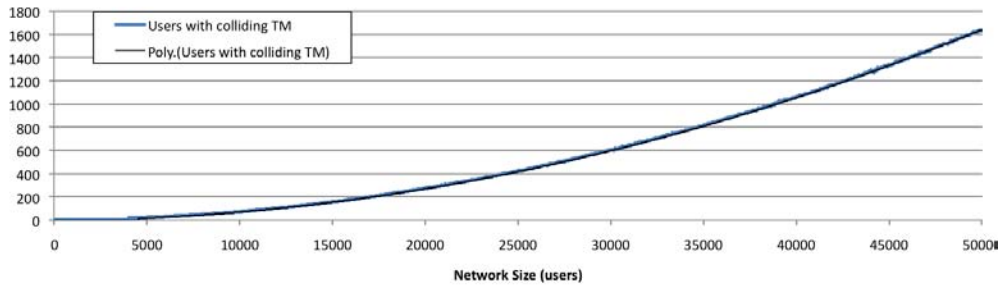
Evaluation Setup

This configuration is likely to reproduce the same cookie timestamp distribution of a company deploying a new OS on its computers network. Therefore, the network using a block of IP addresses might be very large. The size of the bundle in which we are searching for collisions is denoted as n and varies from 50 (the average number of user per bundle according to [27]) to 50,000 users (corresponding to a very large network).

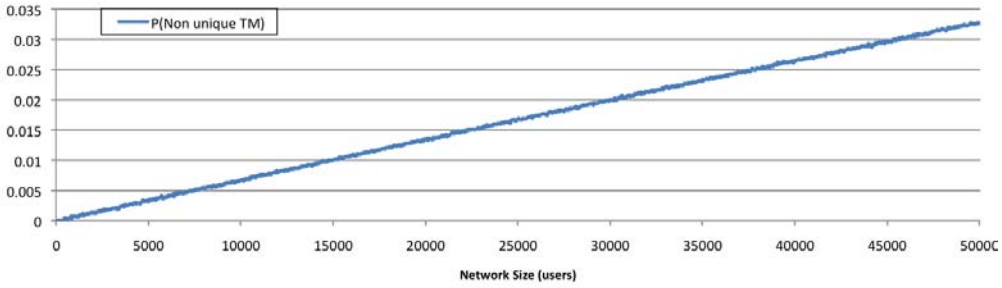
The collision rate is evaluated as follows: n random users are selected in the AOL logs, and the number of them who made their first search simultaneously is observed. For each size of network n, we made 100 evaluations. We then divide the number of collisions by the number of users to obtain the average probability for a user to have the timestamp of at least one other user of the network.

Evaluation Results

The number of collisions as a function of n is depicted in Figure 6(a). This number is zero for n=50 and is negligible for any n smaller than 5000. Therefore, first query timestamp collisions are very unlikely; not only in an average bundle, but also in bundle grouping searches issued from large company networks. The number of collisions then increases quadratically with the network size. There are almost 1,000 collisions in a network of 35,000 computers. We did not evaluate the number of collision in networks containing more than 50,000 users, as it is very unlikely that bundles could contain



(a) Number of collisions



(b) Collision probablity

Figure 6: Estimation of collision rates based on AOL logs

queries issued from such a large number of different computers.

Surprisingly, for n=1000, the collision rate is lower for this set of logs than for the cookies used in the previous evaluation. In fact, this observation may confirm that some cookies used in the previous evaluation belonged to servers that try to translate content via Google translate. Because several requests are issued from the same server simultaneously, the collision rate is increased.

The collision probability, as a function of the network size, is illustrated in Figure 6(b). Even in a very large network, the probability for a user to have a value of TM similar to someone else is less than 4%. Furthermore, as it has been proven by the previous evaluation that even if two users receive the first cookie at the exactly same second, there is a 10% chance that the attached LM values will differ. This result means that average end users cannot expect their searches not to be linkable unless they delete their cookies frequently.

Pseudonymizing Paranoid Users

Finally, even if users renew all of their cookies, they can still be identified unless many users in the same IP address block do the same. To pseudonymize the queries of users deleting their cookies, the attacker could process in three steps:

- An attacker could filter the search queries that are connected to cookies. This approach could be used to reduce the number of unattributed fibers in the bundle to a handful of users.
- The attacker could then use the User Agent string as a quasi-identifier.
- Finally, if some users modified the User Agent the timing and topic analyses can be used to separate the remaining fibers of these users [27].

Consequently, even the searches made by users who delete their cookies could be de-anonymized.

6 Access to the sanitized logs

Although sanitized search logs contain a lot of critical information, one can wonder why these logs are so critical when non-anonymized logs contain obviously identifying information. The answer is twofold.

- These logs are stored for an indefinite period of time. As a result, they could reappear at any given time and could serve as the basis of discrimination for a long time after those searches were issued.
- In addition, these logs are not subject to the same access policies that non-anonymized search logs are. Since sanitized search logs are not supposed to contain personal information, their access is not constrained by Google's privacy policy. These logs can be accessed by Google employees or shared with third parties.

6.1 Internal Employee Policy

In 2007, the congressman Joe Barton asked Google who has access to the retained data. In response, Google said that "access to personal information [is restricted] to Google employees, contractors and agents [...]" (see [7]). However, access to 'non-personal' information, like the sanitized query logs, is not mentioned in this response. In fact, according to Matt Cutts (head of Google's Web spam team) Google's "internal user data access agreement explicitly mentions that Google employees are not allowed to try to access data on any public figure, any employee at a particular company, or any acquaintance" [2]. This discourse emphasizes that the access restriction applies only to logs containing personal information like names or address. Nevertheless, sanitized search logs are certainly not subject to the same policy, and it is unclear whether or not

Google employees have access to these logs. Since these logs have been used in [8], we can assume that some employees at least have access to the sanitized search logs.

The problem becomes more serious when it concerns the intellectual property of corporations whose employees are frequently searching on Google. Indeed, a bundle is very likely to contain only searches issued from this corporation.

6.2 Sharing Logs with Third Parties

Since these logs do not contain Personal Identifiers, they can be assimilated to aggregate data defined by [19], as

> "information that is recorded about users and collected into groups so that it no longer reflects or references an individually identifiable user."

Therefore, according to Google's Privacy policies, these search query logs could be shared with third parties without prior user consent [17]. Until 2010 Google Privacy Policies explicitly mentioned aggregated non-personal, information could be shared with third parties [17]. The October 2010 version [20] makes no mention of Google's policy on third party sharing, one way or the other.

7 Related Work

This section relates previous works that analyzed Google's privacy policy and then overviews log sanitization mechanisms that have been proposed and evaluated.

7.1 Google Policies analysis

Estimating the information that remains in the sanitized search log is particularly critical because the policy that is applied once the logs are sanitized is ambiguous and not mentioned explicitly. Google's privacy policy has already been analyzed and criticized [32, 37]. The lack of clarity and the "loopholes" that this privacy policy contains have been discussed in [32]. [24] discusses Google privacy rhetoric in the media and in its own privacy policy as well as the evolution of this policy over the years. The privacy policy mainly focuses om personal information; this paper is focused on search logs which may not be covered by privacy policies because they do not explicitly contain personal information.

7.2 Query log sanitization

The science of query log sanitization has advanced rapidly since the 2006 AOL search log disclosure. [5] reviews several log privacy-enhancing techniques and algorithms that could be used by search engines: *log deletion*, *queries hashing*, *identifier deletion*, *identifiers hashing*, *query content scrubbing*, *infrequent queries deletion* and *sessions shorten-*

ing. For all these techniques both the log utility preservation and the privacy protections are evaluated. The method developed by Google combines *session shortening* with partial *identifier deletion* and is weaker than each of these techniques applied separately. Unfortunately this specific method is not evaluated.

[26] evaluated the possibility for an attacker to re-identify a user from search history logs. They concluded that scrubbing the search logs might not be enough and that even one day's worth of query could contain enough identifying information to reduce the set of potential identity. In a second analysis, [27] used only the generalized IP address in conjunction with Machine Learning techniques to identify fibers in the anonymized logs. They showed that IP generalization does not provide adequate guarantees. To the best of our knowledge, no experiment has been run to evaluate the risk of re-identifying information in query logs bundles with session identifiers lasting 24 hours. By taking advantage of these quasi-identifiers, an attacker could certainly outperform the techniques proposed in [26, 27].

Some algorithms have been proposed to sanitize search logs that are published or shared. [30] generates a private query click graph containing only queries that are submitted by at least K users (where K is a threshold fixed by the data publisher). The query count is modified by adding a random noise before selecting the queries that are represented in the graph. Authors show that, while the sanitization algorithm provides guarantees similar to differential privacy, the query click graph can still be used to improve query suggestion and spelling correction. [22] propose an algorithm ZEALOUS to build a histogram of query counts that is modified by an additive noise. Similarly to the query click graph[30], ZEALOUS only publishes queries that have been issued by at least θ users. Search logs sanitized with ZEALOUS can still be used for *index caching* and *query substitution*.

8 Discussion

8.1 Remaining issue

In this paper, we focused on search and suggest query logs, because they contain the raw data that Google and other search engines collect. This data that is collected by search engines and used to build accurate profiles, which may be stored separately of the raw data and may not be connected to a personal identifier. Each search reveals one interest, and while the search logs are subject to the retention policy, the interests that search engines infer from searches are not.

Using searches (or interests) in the long run, a search engine may keep track of a user. For instance, the search engine could establish that there is only one user who:

- Is in IP address block 134.167.211.X,
- Searches frequently for terms related to 'ethnography in the middle east,'
- Usually searches from Monday to Thursday, 11PM to 1AM.

At any point in time, the search engine could use this information to reconnect a user's cookie to his profile based on searches made many years before. Indeed, the profile that is inferred from user searches, while not subject to a retention policy, could contain very accurate information about a user's habits and history. This profile may not even reflect a single user but can concern several users sharing either an interest or who are searching from the same office.

8.2 Recommendations

In this paper we demonstrated that log retention policies should not focus only on the identifiers that are anonymized or obfuscated but should cover also the pieces of information that remain in the logs. Although, keeping the User Agent string in anonymized server logs might seem innocuous, it appears that this string makes Google sanitization less effective [9]. Even if these data have some utility, Google should weigh this utility with the utility of other pieces of information, such as the three first bytes of the IP address.

Furthermore, Google announces retention policy updates before their implementations [13, 14]. These announcements are often subject to misinterpretation, and further information is needed to clearly understand the consequences of these updates [34]. Privacy advocates and security experts cannot provide valuable feedback without a better idea of these anonymizing processes.

We recommend, therefore, a mixed strategy: one is for search engines to adopt publicly available, cutting edge sanitization processes, such as those described in Section 6, so the algorithms can be evaluated by the community of peers. Another, as suggested by the EU, is an external audit. The external audit, which may be conducted under non-disclosure constraints, is necessary to see how these algorithms are working on actual data and how they scale with the history logs that are maintained by major search engines. These two complementary approaches will improve and verify the effectiveness of sanitization techniques.

Acknowledgments

We thank the NYU Privacy Research Group for motivating this research and providing useful feedback to an early presentation of results. We thank Alma Whitten for a fruitful and productive conversation, Steve Fienberg and Arvind Narayanan for improving the paper through helpful suggestions on earlier drafts. This research was supported by AFOSR MURI award ONR BAA 07-036 and NSF GENI award CNS-0820795.

References

[1] Article 29 Data Protection Working Party (2007). Data protection in the European Union, May 16th, 2007. `http://ec.europa.eu/justice_home/fsj/privacy/news/docs/pr_google_16_05_07_en.pdf`.

[2] Battelle, J. (2008). Google responds to privacy fears on searchblog.

http://battellemedia.com/archives/2008/02/google_responds_to_privacy_fears_on_searchblog.

[3] Castelluccia, C., Cristofaro, E. D., and Perito, D. (2010). Private information disclosure from web searches (The case of Google web history). *CoRR*, abs/1003.3242.

[4] Chen, C. (2010). Google's approach to privacy. http://googlepublicpolicy.blogspot.com/2009/12/googles-approach-to-privacy.html.

[5] Cooper, A. (2008). A survey of query log privacy-enhancing techniques from a policy perspective. *ACM Trans. Web*, 2: 19:1–19:27. http://doi.acm.org/10.1145/1409220.1409222.

[6] Davidson, A. (2007). Google responds to questions from Congressman Joe Barton. http://searchengineland.com/scoop-google-responds-to-rep-joe-bartons-24-privacy-questions-12999.

[7] — (2008). Google responses to questions from the House Energy and Commerce Committee. http://services.google.com/blog_resources/google_policy_davidson_letter.pdf.

[8] Duebendorfer, T. and Frei, S. (2009). Why silent updates boost security. http://www.techzoom.net/publications/silent-updates/.

[9] Eckersley, P. (2009). How unique is your web browser? Technical report, Electronig Frontier Foundation. https://panopticlick.eff.org/browser-uniqueness.pdf.

[10] Fleischer, P. (2007). Google response to Data Protection Working Party. http://static.googleusercontent.com/external_content/untrusted_dlcp/services.google.com/en/us/blog_resources/Google_response_Working_Party_06_2007.pdf.

[11] — (2008). Privacy at Google. http://www.google.com/events/docs/policyblog_uk_privacy_booklet.pdf.

[12] — (2010). Privacy: A number's game? http://peterfleischer.blogspot.com/2010/09/privacy-numbers-game.html.

[13] Fleischer, P., Horvath, J., and Whitten, A. (2008). Another step to protect user privacy. http://googleblog.blogspot.com/2008/09/another-step-to-protect-user-privacy.html.

[14] Fleischer, P. and Wong, N. (2007). Taking steps to further improve our privacy practices. http://googleblog.blogspot.com/2007/03/taking-steps-to-further-improve-our.html.

[15] Frei, S., Duebendorfer, T., and Plattner, B. (2008). Firefox (in) security update dynamics exposed. *SIGCOMM Comput. Commun. Rev.*, 39: 16–22. http://doi.acm.org/10.1145/1496091.1496094.

[16] Google (2007). Google log retention policy FAQ. http://publicintelligence.net/google-log-retention-policy-faq/.

[17] — (2009). Google privacy policy 2009. http://www.google.com/privacy_archive_20090127.html.

[18] — (2010). Google Chrome and privacy. http://static.googleusercontent.com/external_content/untrusted_dlcp/www.google.com/en/us/intl/en/landing/chrome/google-chrome-privacy-whitepaper.pdf.

[19] — (2010). Google privacy FAQ. http://www.google.com/intl/en/privacy_faq.html.

[20] — (2010). Google privacy policy 2010. http://www.google.com/intl/en/privacypolicy.html.

[21] Google-Privacy-Channel (2007). Google search privacy: Plain and simple. http://www.youtube.com/watch?v=kLgJYBRzUXY.

[22] Götz, M., Machanavajjhala, A., Wang, G., Xiao, X., and Gehrke, J. (2009). Privacy in search logs. *CoRR*, abs/0904.0682.

[23] Hitwise (2010). Top search engines in 2010. http://www.hitwise.com/us/datacenter/main/dashboard-10133.html.

[24] Hoofnagle, C. J. (2009). Beyond Google and Evil: How Policy Makers, Journalists and Consumers Should Talk Differently About Google and Privacy. *First Monday*, 14(4–6).

[25] Hözle, U. (2010). Update to Google Suggest. http://googleblog.blogspot.com/2008/09/update-to-google-suggest.html.

[26] Jones, R., Kumar, R., Pang, B., and Tomkins, A. (2007). "I know what you did last summer": Query logs and user privacy. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*, CIKM '07, 909–914. ACM. http://doi.acm.org/10.1145/1321440.1321573.

[27] — (2008). Vanity fair: Privacy in query log bundles. In *Proceeding of the 17th ACM Conference on information and Knowledge Management*. New York, NY, USA: ACM.

[28] Keize, G. (2008). Google bends to Chrome privacy criticism. http://www.computerworld.com/s/article/9114369/Google_bends_to_Chrome_privacy_criticism?taxonomyId=84&pageNumber=2.

[29] Kohnstamm, J. (2010). Letter from the Article 29 Working Party addressed to search engine operators. http://ec.europa.eu/justice_home/fsj/privacy/docs/wpdocs/others/2010_05_26_letter_wp_google.pdf.

[30] Korolova, A., Kenthapadi, K., Mishra, N., and Ntoulas, A. (2009). Releasing search queries and clicks privately. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, 171–180. New York, NY, USA: ACM. http://doi.acm.org/10.1145/1526709.1526733.

[31] Lynch, B. (2009). Comparing search data retention policies of major search engines before the EU. http://blogs.technet.com/b/microsoft_on_the_issues/archive/2009/02/10/comparing-search-data-retention-policies-of-major-search-engines-before-the-eu.aspx.

[32] Meuli, G. and Finn, C. (2007). Google: Trust, choice, and privacy. http://www.ethicapublishing.com/ethical/3CH15.pdf.

[33] Protalinski, E. (2010). Microsoft: Google Chrome doesn't respect your privacy. http://arstechnica.com/microsoft/news/2010/03/microsoft-google-chrome-doesn-your-privacy-microsoft-google-chrome-doesnt-respect-your-privacy.ars.

[34] Soghoian, C. (2008). Debunking Google's log anonymization propaganda. http://news.cnet.com/8301-13739_3-10038963-46.html.

[35] Sullivan, D. (2008). Anonymizing Google's server log data—How's it going? http://searchengineland.com/anonymizing-googles-server-log-data-hows-it-going-15036.

[36] Sweeney, L. and Sweeney, L. (2002). Achieving k-anonymity privacy protection using generalization and suppression. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5): 571–588.

[37] Tene, O. (2007). What Google Knows: Privacy and Internet Search Engines. *SSRN eLibrary*. http://ssrn.com/paper=1021490.