

BiscuitSpy: Profiling web users via HTTP cookies

Marcela Melara, Jenny Guo

Abstract

ABSTRACT HERE

1 Introduction

Example of incorporation citations [?].

2 Background

Global cookies, such as DoubleClick's `id` cookie, have a globally unique value, *i.e.*, its value remains the same across different visited websites in each *session* (**NEEDS VERIFICATION: per session, or per lifetime**). Ad companies use these cookies to track user through their web and collect information about their browsing habits and personal preferences.

Local cookies, on the other hand, such as the Google Analytics `__utma` cookie, have locally unique values. This means that each unique domain that utilizes such a cookie to track an individual user will have a unique user identifier for the entire *lifetime* of the cookie (**NEEDS VERIFICATION: per session, or per lifetime**). These kinds of cookies are used by domains to keep track of their own users and their browsing habits pertaining to their website.

2.1 Cookies used by Google

As Google cookies are one of the most pervasive cookies on the Internet we start by giving an overview of the different types of cookies that Google uses. They can be divided into the following categories:

- Preferences
- Security
- Processes
- Advertising
- Session State
- Analytics

Preference cookies allow Google to keep track of user preferences in order to provide a more personalized browsing experience. Information stored include default language, user region, number of search results to display per page, text size and font and other preferences. Notably through the use of these cookies, Google does not require the user to be signed in to provide the targeted Google website. The PREF cookie is the most dominant cookie for storing user preferences. It also includes a timestamp of the most recent user preference change. This unique information can be used to pinpoint individual users due to the unlikelihood that two users have the same timestamp.

Security cookies are used to authenticate users and prevent fraudulent use of login credentials. They store the Google account ID and most recent sign-in time in an encrypted and signed string. Security cookie names end with SID, such as SID, HSID, SAPISID and APISID. The APISID cookie and SAPISID cookie both contain two encrypted strings, where the second string is *unique across all Google websites and per cookie lifetime*. Therefore, despite the encryption, it is still possible to identify individual users and profile what Google services the user is visiting.

Process cookies support the display and function of more complex and dynamic websites. Google states that these cookies are necessary for the proper functioning of some websites. (**REFERENCE**) For example blocking an 'lbc's' cookie would prevent Google docs from opening documents correctly. BiscuitSpy does not include this cookie in its profiling.

Advertising cookies are used to determine which ads to display to the user and for tracking the user's ad clicking behavior through a complex network of publishers, advertisers and website operators. The most commonly seen advertising cookie is the 'id' cookie from the domain ad.doubleclick.net. The second field of the 'id' cookie contains several numbers that remain the same during multiple visits of the same website, but change their values either when the user has clicked on an advertisement on the website, or when the website displays a different ad. It can thus be inferred that the numbers encode the ads to display and whether the user has clicked on them.

Session state cookies collect information about how users interact with a website and keep information of the previous sessions of a website. Youtube session cookies for example store a list of most recent videos watched in that browser. Session state cookies are also used to measure the effectiveness of affiliate advertising. It is difficult to determine the exact names of session cookies and the meaning of their values, therefore BiscuitSpy currently does not leverage session cookies for user profiling.

Analytics cookies represent the largest group of Google cookies and are mainly used by BiscuitSpy to gather user profile information. While the previous cookies all belong to a Google domain, analytics cookies do not have that restriction and can be found on all websites that use Google Analytics (GA) and are stored under the current website's domain. The five main cookies set by GA are _utma, _utmb, _utmc, _utmv and _utmz.

As an example, Figure 1 shows the individual components of the _utma cookie. The timestamp information of the initial, most recent and current visit in combination with the number of visits to the website allow for an accurate reconstruction of user's browsing behavior.

```
utma=123456789.123456789.1234567890.1234567890.1234567980.1
```

Cookie Field	Description
Domain Hash	The first number is the domain hash. This is set by all cookies from this domain.
Visitor ID	The second number is a random "unique ID".
Initial visit	The third number is the unix time stamp for the initial visit and is set as soon as you enter the site.
Previous Session	The third number is the unix time stamp for the previous session.
Current Session	The fourth nubmer is the unix time stamp for the current session.
Session number	Contains the number of visits to this website since cookie creation.

Figure 1: Google Analytics UTMA cookie

3 Profiler Design

To demonstrate the feasibility and ease of identifying and distinguishing individual users on a WiFi network based solely on observed advertising and analytics HTTP cookies, BiscuitSpy is comprised of three main components:

1. Capturer: A packet sniffing tool to capture any HTTP packet travelling through the network.
2. CookieBowl: A utility for collecting cookies and extracting the relevant cookies based on pre-defined criteria.
3. Profiler: A tool which parses the information from the extracted cookies to aggregate this data and build a user profile.

The Capturer monitors the WiFi network for any HTTP request or response packets, and filters out any such packets which do not contain cookie the "Cookie" or "Set-Cookie" headers. Appropriate packets are then sent to the CookieBowl; for each captured packet, the CookieBowl first extracts the names and values of all the cookies contained within into a map data structure. Next, the CookieBowl searches for a pre-defined set of advertising and analytics cookies, the *profiling cookies*, among all collected cookies, and separates out these relevant cookies.

The Profiler, takes the map of found profiling cookies and parses the data found in each cookie's value. As discussed in Section 2, BiscuitSpy uses two types of advertising/analytics cookies to build a user profile: (1) *global cookies*, and (2) *local cookies*. Since the value of global cookies does not change across different visited websites, BiscuitSpy can cross-reference each appearance of such cookies allowing it to link two independent HTTP requests captured by the packet sniffer during a single BiscuitSpy session. At the same time, local cookies, which contain a per-domain unique user identifier for the entire *lifetime* of the cookie (**NEEDS VERIFICATION: per session, or per lifetime**) allow BiscuitSpy to infer crucial user browsing behavior. By combining common global and local ad and analytics cookies, we are able to track individual users throughout a

single browsing session using global cookies while gathering browsing behavior information via the associated local cookies. Figure 2 shows the data flow between the various components of BiscuitSpy.

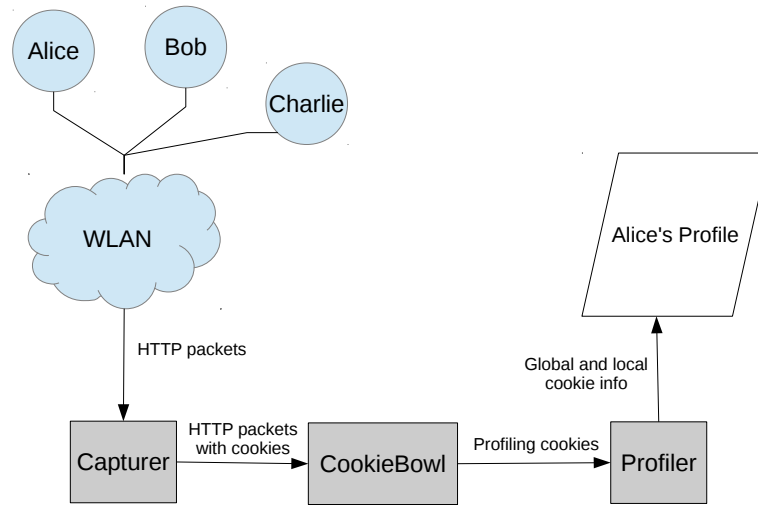


Figure 2: The three main components of the BiscuitSpy tool and the data flow between them.

For each individual user identified by the Profiler, it stores the aggregated browsing behavior in a file for later offline analysis. As cookies mostly contain encoded information, the Profiler converts this data to a human-readable format to facilitate the creation and reading of the profile. One important note to make is that BiscuitSpy is limited by the type of data stored in the pre-defined profiling cookies. Therefore, an analyst may not be able to learn personal information like a user’s name or date of birth from her BiscuitSpy profile unless one of the profiling cookies contains an identifier, *e.g.*, an email address, linkable to a real-world identity. Nonetheless, even without such data, a user browsing profile could still be used to infer a user’s habits or current issues in her life, as well as re-identify her during a later browsing session should she visit some or all of the same websites recorded in her initial profile.

4 Implementation

We have implemented a basic command line-based BiscuitSpy prototype¹ in Java (**ADD SLOC count?**) developed for Linux Ubuntu 12.04 LTS. Our code is divided into three main packages: `cookies` which contains the `CookieBowl` as well as a general `Cookie` class, `cookies.definitions` which contains subclasses of `Cookie`, one for each profiling cookie, and `profiler` which contains both the `Profiler`, the `Capturer`, and a general `Profile` class. Because the `Capturer` uses the `jnetpcap-1.3.0` library, a full Java implementation of the `pcap` API for capturing network traffic, we have bundled its source with the `Profiler`’s to facilitate communication between the two classes.

The `Profiler` keeps a list of each `Profile` so that it can search through all `Profiles` seen so far and potentially annotate these with new information as it discovers cookies that

¹The source code can be found at <http://github.com/naturegirl/BiscuitSpy.git>.

belong to the same user. Before ending a BiscuitSpy session, the analyst can choose to save some or all profiles to separate files for later analysis.

5 Evaluation

We tested our BiscuitSpy prototype on the an open home wireless LAN with MAC Address filtering as well as on the open Princeton University WiFi network, which also has MAC Address filtering. Due to factory configuration issues, we were never able to capture useful packets from devices on either network. However, we deem our current prototype sufficient for demonstrating the feasibility of profiling a user based on HTTP cookies.

Our results show that almost all cookies contain a unique browser identifier which can be used to pinpoint individual users. Furthermore we identified three main types of personal information that can be captured through cookies:

1. Browsing habits are leaked through timestamps
2. Location information is leaked in cookies
3. Username and e-mail are leaked in cookies

5.1 Leaking of browsing habits

The `_utma` cookie is one of the most prevalent cookies and is used by BiscuitSpy to create a profile of the user's browsing habits. For any website the user visits that contains this cookie BiscuitSpy can determine the initial visit time to that website, the most recent visit time, the current visit time and the number of visits. Especially the initial visit combined with the number of visits allows us to infer how long the user has at least used the current browser and whether he regularly visits the website. The example in figure 3 is taken from the author's `_utma` cookie of www.github.com.

```
__utma=1.348058162.1366417185.1370673194.1389377452.11  
first Visit: Sat, 20 Apr 2013 00:19:45 GMT  
previous Visit: Sat, 08 Jun 2013 06:33:14 GMT  
current Visit: Fri, 10 Jan 2014 18:10:52 GMT  
count: 11
```

Figure 3: utma cookie with parsed timestamp information

5.2 Leaking of location information

We also discovered that certain cookies leak very detailed geographical information. For example <http://www.washingtonpost.com/> contains a `rp1d1` cookie that tells us that the user is accessing the website through a `princeton.edu` network, that he is located in Princeton, New Jersey and even passes the longitude and latitude of his current location in plaintext!

```
rpld1=0:princeton.edu|20:usa|21:nj|22:princeton|  
23:40.348999|24:-74.658997|
```

Figure 4: rpld1 cookie from washingtonpost.com leaking location information

5.3 Leaking of username and e-mail

While examining cookie data captured from BiscuitSpy we came across a particularly disturbing cookie from wordpress.com called `wpc_wpc`

```
|account=xuezhongwen&  
avatar=https%3A%2F%2F1.gravatar.com%2Favatar%2F420824352fe495c1b3b1dd0  
bc157def4%3Fs%3D25%26amp%3Bd%3Dhttps%253A%252F%252Fs2.wp.com%252Fwp-con  
tent%252Fmu-plugins%252Fhighlander-comments%252Fimages%252Fwplogo.png&  
email=nature_girl473%40yahoo.de&  
link=http%3A%2F%2Fgravatar.com%2Fxuezhongwen&  
name=naturegirl&  
uid=1757819&  
access_token=1d31013fd9a4613cfb7975c98553c4f762f32750";
```

Figure 5: wpc_wpc cookie from wordpress.com leaking username and e-mail

This cookie transmitted the username and e-mail in plaintext and was present on all websites that are based on wordpress.com. For example a visit to the website <http://www.hustlermoneyblog.com> transmits the cookie as seen in figure 5 which contains the user's email address, username and user id.

The above three types of user data leaks allow BiscuitSpy to create a user profile that contains his browsing habits, location information and username and e-mail. In addition BiscuitSpy will also collect unique browser identifier cookies to be able to cross reference visits to different websites and recognize users in a new browsing session.

5.4 Cookies on Alexa Top 10 websites

In addition to determining particularly critical personal information leaks that can be used to profile a user, we also looked at the top 10 websites as determined by Alexa and examined their cookie usage. We only focused on the English websites out of the top 10 websites (excluding the 3 Chinese websites baidu.com, qq.com and taobao.com) and were specifically interested in what fraction of them uses the cookies as described in paragraph 2 and used by BiscuitSpy for profiling.

In general, we see from Table 1 that these websites tend to rely on their own cookies. Only two websites have the doubleclick 'id' cookie and the Google PREF cookie as third party cookies. It is also notable that wikipedia uses the fewest cookies, with 0 cookies on their start page <http://www.wikipedia.org> and only 4 cookies on their main page http://en.wikipedia.org/wiki/Main_Page

Website	own domain cookies	third party cookies
google.com	8 cookies including PREF and SID	none
facebook.com	12 cookies including 'datr'	none
youtube.com	9 cookies	doubleclick 'id', google PREF and SID
yahoo.com	19 cookies	none
wikipedia.org	4 cookies	none
amazon.com	9 cookies including apn-user-id	doubleclick 'id'
live.com	9 cookies	none

Table 1: Cookie usage on English websites from Alexa Top 10 websites

6 Technical and Legal Context with Surveillance

7 Related Work

8 Future Work

Example of incorporation citations [?].

9 Conclusions