BiscuitSpy: Profiling web users via HTTP cookies

Marcela Melara, Jenny Guo

1 Introduction

HTTP cookies were originally introduced as a mechanism to make websites stateful [7]. Websites use these small pieces of data to keep track of a user's browsing activity, such as whether she is logged in, or which items she has added to her shopping cart, and they store these cookies on the client-side in the user's web browser [1, 7]. When the user loads the website, the browser will send the appropriate stored cookies back to the web server to return to the last recorded state. However, HTTP cookies are not only being used to improve the user online experience. Because cookies allow websites to remember a given user's browsing activity or preferences on that site, online advertising companies have found a way to leverage HTTP cookies and take advantage of the vast amounts of available user information online. At the same time, website publishers themselves now often employ *analytics cookies*, special HTTP cookies used collect statistics about their users and the usage of their website.

In addition to the cookies required for the functionality of a website, third-party advertising companies include *tracking cookies* and first-party website publishers add analytics cookies when a user visits this site [1]. Such cookies have been raising concerns about users' privacy online for years because companies can identify and target consumers, and compile a vast browsing history for users and learn about their personal preferences and habits [5, 8]. Moreover, researchers have found that cookies can leak personal information belonging to any user to third-party sites via online social networks allowing these advertisers to learn the viewing habits of some user, and associate these viewing habits with a specific person [4]. In particular, Krishnamurthy *et al.* [4] determined that third-party advertising companies are able to obtain such information through a combination of HTTP header information and cookies.

The Firesheep project[2] demonstrated the ease with which an adversary sniffing packets on an unencrypted network can hijack others' browsing sessions by obtaining authentication cookies [6]; many affected sites responded by implementing an HTTPS-only policy. However, recent revelations about the National Security Agency leveraging Google advertising cookies to pinpoint targets for hacking and surveillance (*e.g.,* [3, 8]) serve as a reminder that users are still vulnerable to third-party tracking based on HTTP cookies. Taking a similar approach to Firesheep, our work aims at raising awareness about user profiling based on HTTP cookies used for advertising and analytics. To this end, we present BiscuitSpy, a tool which allows a user to observe web traffic in an unencrypted network, collect HTTP cookie information, and aggregate these data to build a browsing profile of a user. BiscuitSpy leverages existing first- and third-party cookies to define a set of *profiling cookies*, common advertising and analytics cookies found on most popular websites, and filters the captured web traffic for these pre-defined cookies. To create the user browsing profile, BiscuitSpy extracts the information contained in the observed profiling cookies and stores it in files for later offline analysis. We have implemented a basic BiscuitSpy prototype and we have conducted a small-scale measurement study to determine to most common advertising and analytics cookies suitable for profiling.

This paper is organized as follows. In Section 2 we provide relevant background information on cookies and web tracking. Section 3 details BiscuitSpy's system design and Section 4 describes our prototype implementation; we evaluate BiscuitSpy in Section 5. We discuss BiscuitSpy in the context of surveillance technology and law in Section **??** Section 6 describes some related work, and we discuss directions for future work in Section 7. We conclude in Section 8.

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:

2.1.1 Preferences

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.

2.1.2 Security

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.

2.1.3 Processes

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 blooking an ’lbcs’ cookie would prevent Google docs from opening documents correctly. BiscuitSpy does not include this cookie in its profiling.

2.1.4 Advertising

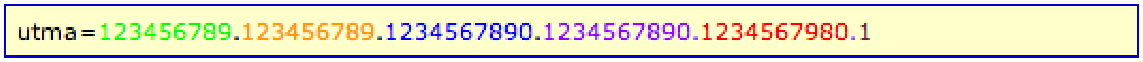
Advertising cookies are used to determine which ads to display to the user and for tracking the user’s ad clicking behavior through a complex network of publishers, advertisers and website operators. The most commonly seen advertising cookie is the ’id’ cookie from the domain ad.doubleclick.net. The 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.

2.1.5 Session State

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.

2.1.6 Analytics

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.



```
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

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.

For the BiscuitSpy implementation we leveraged preference cookies, advertising cookies and analytics cookies. We did not filter out process cookies and session state cookies, since we were unable to identify the specific cookie names and individual cookie field meanings.

2.2 Common third-party cookies

While Google cookies are the most common cookies found on websites, we also identified several other common third-party cookies from amazon, facebook and twitter.

A common Amazon third party cookie is the `apn-user-id` cookie. This cookie contains the user-id as shared by the amazon partner's network. The most common Facebook third party cookie is the `datr` cookie. This cookie encodes a user id and remains the same across websites and browsing sessions. Twitter's third party cookie is the `guest_id` cookie which contains a version number and an encoded user id. It is interesting to see how these other non-google cookies become more and more prevalent on the web and challenge Google's position as the dominent user and ad tracking cookie provider.

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.

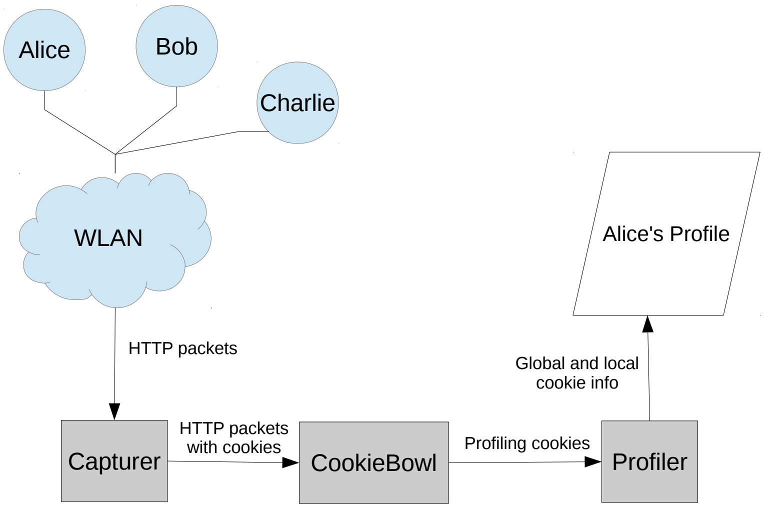


Figure 2: The three main components of the BiscuitSpy tool and the data flow between them.

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 predefined 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 prototype[1] 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.

[1] The source code can be found at `http://github.com/naturegirl/BiscuitSpy.git`.

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 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

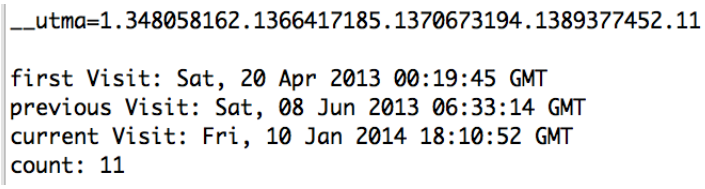
We tested our BiscuitSpy prototype on the an open home wireless LAN with MAC Address filtering as well as on the open Princeton University WiFi network, which also has MAC Address filtering. Due to factory configuration issues, we were never able to capture useful packets from devices on either network. 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 intial 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.



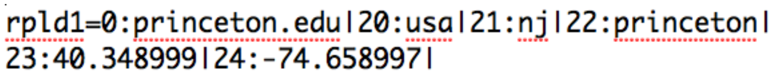
```
__utma=1.348058162.1366417185.1370673194.1389377452.11

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 rpld1 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!

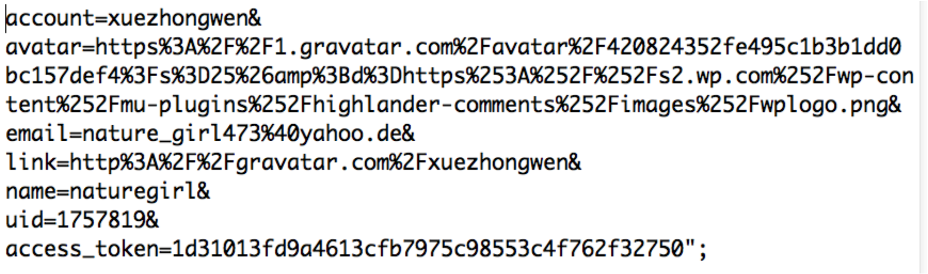


```
rpld1=0:princeton.edu|20:usa|21:nj|22:princeton|
23:40.348999|24:-74.658997|
```

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 user data 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 Related Work

7 Future Work

Example of incorporation citations [**?**].

Currently BiscuitSpy only sniffs http packets from the user's own computer. In the future it would be important to extend BiscuitSpy to also be able to capture packets from other computers in the open wireless network. This would allow the profiling of multiple users at the same time to see how accurate BiscuitSpy can associate individual cookies with the correct user.

In addition BiscuitSpy should include the remaining Google cookies. In the present implementation BiscuitSpy heavily utilizes analytics cookies combined with preference, security and ad tracking cookies. However session state cookies as documented in section 2 also contain sensitive user browsing information. After identifying specific session state cookies and the meanings of the individual cookie fields, BiscuitSpy should include those cookie definitions .

A last area of future work is the GUI of BiscuitSpy. Currently BiscuitSpy has a console interface and supports writing output to the screen and to a file for further analysis. For easier usage and a better overview of the user profile a GUI should be added to it.

8 Conclusion

BiscuitSpy was built as a proof-of-concept software for profiling web users based on http cookies. It offers functionality to capture packets on an open wireless network and filter out http headers which transmit cookie data. The captured cookie data is combined with existing cookie definitions to extract relevant user information. In our work we showed that captured cookies leak various personal information, such as browsing behavior, location information, username and e-mail address and unique brower identifiers. By aggregating the cookie information from multiple sources, BiscuitSpy can built up a comprehensive user profile.

References

[1] J. Barnes. Internet users' privacy concerns may mean cookies start to crumble. http://www.theguardian.com/technology/blog/2013/may/24/

internet-privacy-cookies-firefox, May 2013.

[2] Butler, Eric. Firesheep. https://github.com/codebutler/firesheep.

[3] Kamdar, Adi and Reitman, Rainey and Schoen, Seth. NSA Turns Cookies (And More) Into Surveillance Beacons. https://www.eff.org/deeplinks/2013/12/nsa-turns-cookies-and-more-surveillance-beacons, Dec 2013.

[4] B. Krishnamurthy and C. E. Wills. On the leakage of personally identifiable information via online social networks. *SIGCOMM Comput. Commun. Rev.*, 40(1):112–117, Jan. 2010. ISSN 0146-4833. doi: 10.1145/1672308.1672328. URL http://doi.acm.org/10.1145/1672308.1672328.

[5] E. Lee. Sen. Rockefeller: Get Ready for a Real Do-Not-Track Bill for Online Advertising. http://adage.com/article/digital/sen-rockefeller-ready-a-real-track-bill/227426/, May 2011.

[6] S. Machlis. How to Hijack Facebook Using Firesheep. http://www.pcworld.com/article/209333/how_to_hijack_facebook_using_firesheep.html, Oct 2010.

[7] Microsoft Developer Network. HTTP Cookies. http://msdn.microsoft.com/en-us/library/windows/desktop/aa384321%28v=vs.85%29.aspx, Retrieved Jan. 2014.

[8] A. Soltani, A. Peterson, and B. Gellman. NSA uses Google cookies to pinpoint targets for hacking. http://www.washingtonpost.com/blogs/the-switch/wp/2013/12/10/nsa-uses-google-cookies-to-pinpoint-targets-for-hacking/, Dec 2013.