This article is downloaded from

http://researchoutput.csu.edu.au

It is the paper published as:

Authors: Rath Kanha Sar, and Yeslam Al-Saggaf

Title: Social networking sites’ tracking of unintentionally shared information

Journal Title: First Monday     ISSN: 1396-0466

Year: 2013     Volume: 18     Issue: 6     Pages: 17pgs

Abstract: Various pieces of information are being shared online while users browse on the Internet. Users’ information is being shared or leaked from first party or visited sites to third party sites (such as advertisers) in a number of ways including in the HTTP headers. Within this study, we analysed HTTP headers resulting from the first author’s browsing activities and reported on the types of information being leaked or shared, and to whom. We observed that within just a single browsing session among some social network sites (SNSs) and sites that are not SNSs (non-SNSs), both user’s identifiable and non-identifiable information are being leaked or shared to various third party sites and also propagated to more than just one level of third party sites. In addition, we also discovered that SNSs such as Facebook, Twitter, and Google Plus, are able to track users’ browsing activities not only within SNSs but also beyond the SNS boundaries, particularly among web sites that embed SNS widgets (Facebook’s Like button, Twitter’s Tweet button, and Google’s Plus One button).

DOI/URL:  http://dx.doi.org/10.5210%2Ffm.v18i6.4349

http://researchoutput.csu.edu.au/R/-?func=dbin-jump-full&amp;object_id=47059&amp;local_base=GEN01-CSU01

Author Address: yalsaggaf@csu.edu.au

CRO Number: 47059
Propagation of unintentionally shared information and online tracking

Abstract:
Various pieces of information are being shared online while users browse on the Internet. Users’ information is being shared or leaked from first party or visited sites to third party sites (such as advertisers) in a number of ways including in the HTTP headers. Within this study, we analysed HTTP headers resulting from the first author’s browsing activities and reported on the types of information being leaked or shared, and to whom. We observed that within just a single browsing session among some social network sites (SNSs) and sites that are not SNSs (non-SNSs), both user’s identifiable and non-identifiable information are being leaked or shared to various third party sites and also propagated to more than just one level of third party sites. In addition, we also discovered that SNSs such as Facebook, Twitter, and Google Plus, are able to track users’ browsing activities not only within SNSs but also beyond the SNS boundaries, particularly among web sites that embed SNS widgets (Facebook’s Like button, Twitter’s Tweet button, and Google’s Plus One button).

Contents

1. Introduction
Concerns about the impact that technologies have on informational privacy arise when different types of information about individuals can be rapidly collected, in large volumes, without users being aware of the collection, or how their information will be used, or the duration information will be kept (Tavani, 1999, 2011). Large volumes of information are available in topical databases or diffused online. We categorised information diffused online into two types: the “intentionally” and the “unintentionally” shared information.

The term “intentionally shared information” refers to information that is voluntarily supplied by users to an intended recipient group. For example, people share their opinions in web forums and personal blogs or on social network sites (SNSs) with an expectation or awareness that their posts could be seen by other online users. SNSs could be the biggest space where large volumes of personal information are being voluntarily shared by SNS users owing to the fact that there are at least one billion users active on SNS like Facebook (ABCNews, 2012), which accounts for one seventh of the world’s population. SNS privacy settings, for example, are seen to be one of the many factors that make the users' information available to the public beyond their expected audience (Gross & Acquisti, 2005; Krishnamurthy & Wills, 2008). Some users do not change their privacy settings because they are not aware that they can do it, or because they do not have the technical knowledge to do it (boyd & Hargittai, 2010), while others are aware of the privacy settings and do care about limiting access to their profiles (Strater & Lipford, 2008; Young, 2009; Madden, 2012). Users may open themselves to risks such as embarrassment, stalking, identity theft, phishing attack as well as scamming, which could harm them physically and mentally (Gross & Acquisti, 2005; boyd & Heer, 2006).

On the other hand, the term “unintentionally shared information” refers to information about a user which can be gathered or leaked without the user’s intention to reveal, or without the user’s awareness that the information is being shared. This could be information about online browsing activities. For example, John is searching for allergy tablets and nasal spray on a pharmacy website.
He might not be aware that his searches can be collected by not only the pharmacy website, which is the first party site, but also by various third party sites such as advertisers, and he did not intend to disclose these types of information to those third party sites. The term first party sites refer to sites that are directly or intentionally accessed by users, whereas third party sites are those which are not directly requested by users. Third party sites can be data aggregators or advertisers who display targeted advertising content on the first party web pages. Users’ online movements or browsing behaviours can be tracked or recorded by various technologies including HTTP cookies (Angwin, 2010; Tene & Polonetsky, 2012). HTTP cookies are seen to be troublesome for privacy because they allow users’ movements to be tracked from site to site.

HTTP, known as Hypertext Transfer Protocol, is used to communicate between the browser, or any intermediate machines, and the web servers (Comer, 2000). Since HTTP is a stateless protocol, each request or response is treated independently. So in order to remember the state, it uses a small text file called an HTTP cookie which is stored by the browser at the user’s machine (Kristol, 2001). Cookies were first developed in 1994 for the purpose of assisting users in online shopping by serving as a virtual shopping trolley (Hormozi, 2005). Cookies were not intended to be used as a spying mechanism, rather for the purpose of informing the web server that the same user has returned. HTTP cookies are not programs, nor can they be accessed at anytime by the website, however, they can be used by marketers and web developers to collect personal information and to track users’ visits and browsing habits.

Previous studies have shed light on the current practice of information leakage or gathering from different perspectives such as via privacy settings of SNSs (Gross & Acquisti, 2005; Krishnamurthy & Wills, 2008), via flash cookies (Soltani, Canty, Mayo, Thomas, & Hoofnagle, 2009; Ayenson, Wambach, Soltani, Good, & Hoofnagle, 2011) and via HTTP headers (Krishnamurthy & Wills, 2006; Krishnamurthy, Malandrino, & Wills, 2007; Krishnamurthy & Wills, 2008, 2009; Soltani et al., 2009; Krishnamurthy & Wills, 2010a, 2010b; Mayer, 2011; Krishnamurthy & Naryshkin, 2011). Those studies investigated large numbers of sites, including social network sites (SNSs) and other sites that are not SNS (or non-SNSs). Those results, however, do not really reflect the browsing habits of users in real life, because users tend to have a combination of browsing among both SNS and non-SNS (Purcell, 2011). They are not likely to browse large numbers of SNS alone (e.g. 10 SNS at a time), and some of them may have their browsers in a private browsing mode (Bursztein, 2012) which means they regularly clear search histories and cookies from their browsers, and they are likely to have a new online browsing session every time they close and restart their browser. In addition, it is important to examine the information sharing among both SNS and non-SNS because while the information leaked from SNSs tends to be more personal and identifiable to a specific user (e.g. name, e-mail address and postcode) (Krishnamurthy & Wills, 2008; Soltani et al., 2009), the combination of this identifiable information and the browsing behaviours among non-SNS could reveal so much about a person’s life.

The question remains, what are the ramifications of information leakages from one browsing session among both SNS and non-SNS? Based on the literature and the rationale above, we intend to investigate the ramifications of the unintentionally shared information that is being diffused in the HTTP headers from the perspective of a user who usually browses both SNS and non-SNS while also regularly clearing the search histories and cookies. We examine the HTTP headers resulting
from the first author’s browsing activities while also reporting on the types of the shared information, and to whom they are being shared.

This study is similar in some ways to the ones conducted by Krishnamurthy and Wills (2010a, 2010b); Mayer (2011); Krishnamurthy and Naryshkin (2011), for it investigated information leakages in the HTTP headers, and it is also similar to a study by Dwyer (2009) and Wongyai and Charoeunwatana (2012) for it employed a small case study which used Wireshark as a tool. However, the differences lie in the choice of number and categories of sites and the choice of online activities. This study did not investigate large numbers of either SNS or non-SNS alone (Krishnamurthy & Wills, 2010a, 2010b; Mayer, 2011; Krishnamurthy & Naryshkin, 2011), nor did it focus on only one organisation (Dwyer, 2009) or site (Wongyai & Charoeunwatana, 2012). Rather, this study examined a small number of sites frequently visited by most people while employing different sets of browsing activities and Google search trends common among most people (see justification in Section 2.1). Several browsing sessions were conducted and each browsing session consisted of a set of browsing activities as described in Section 2.2.

### 2. Methods and data collection

#### 2.1: Online activities, sites, and online search trends for study

First we decided on the set of online activities to be performed based on a report by Pew Internet (Purcell, 2011), then we selected the sites associated with the chosen activities based on the ranking in Alexa (www.alexa.com). In terms of online search, we relied on Google statistics (Google, 2012b) for the top search trends. The top or most popular online activities surveyed by Pew Internet (Purcell, 2011) were checking e-mails, using SNS, doing online shopping, reading online news articles, and performing online searches (e.g. by using Google search engines). After choosing top visited sites (ranked by Alexa) combined with popular online activities (Purcell, 2011) and Google search trends (Google, 2012b), we developed a list of sites as summarised in Table 1.

<table>
<thead>
<tr>
<th>E-mail</th>
<th>SNS</th>
<th>Online shop</th>
<th>News</th>
<th>Google search</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yahoo</td>
<td>Facebook</td>
<td>Ebay</td>
<td>9 MSN news</td>
<td>Lyrics007</td>
</tr>
<tr>
<td>Gmail</td>
<td>Twitter</td>
<td></td>
<td>ABC news</td>
<td>Taste.com.au</td>
</tr>
<tr>
<td></td>
<td>LinkedIn</td>
<td></td>
<td></td>
<td>Weatherzone</td>
</tr>
<tr>
<td></td>
<td>Youtube</td>
<td></td>
<td></td>
<td>Wikipedia</td>
</tr>
</tbody>
</table>

Table 1: Online activities and sites chosen for the studies

Wireshark was our chosen data collection tool. It is a network protocol analyser that captures and displays all HTTP traffic among the network communication using HTTP protocol, such as the communication between the browser or application and the requested sites or servers (Figure 1). The rationale for choosing Wireshark is that we could triangulate our data by conducting the experiments across different Operating Systems (OS) and different network environments (with and without proxy settings), and unlike Fiddler, Wireshark is more versatile and can be installed on any OS. It can also be used to record the activities of any application or browser running on the device. Each recorded
performance can be saved or exported into different file formats for later analysis. The only drawback is that encrypted information or information transmitted over SSL or HTTPS packets are not observable. However, as our aim was to analyse HTTP packets, this encrypted information has no effect on our study.

Figure 1: File captured by Wireshark

Section 2.2: Data collection and analysis

Each experiment involved the first author performing a set of browsing activities among the selected sites (as summarised in Table 1) while having her activities recorded by Wireshark. The experiments were conducted on Windows, Linux and Mac OS machines using Firefox browser, with and without the ad-block Firefox extension, over two separate networks (with and without a proxy server). Users are often encouraged to create accounts for many categories of sites they visit. However, for this study, the first author already owns accounts for sites under investigation, so our work did not involve investigating the possible information leakage/sharing during the sign up process.

Examples of first author’s actions included but was not limited to the following:

- E-mail: signing in, checking, reading and sending e-mails.
- SNS: signing in, checking her own and her friends’ profile, checking messages, where feasible, playing the third party applications and clicking on the advertisements.
- Online shops: signing in, searching for and occasionally purchasing a few items (e.g. hair accessories).
- News: browsing to different types of articles (e.g technology, health, or national news).
- Online search: checking the daily weather forecast for her current location, searching for some general knowledge about a specific topic (e.g. hay fever).
Each experiment or browsing session included the following procedures:
• Making notes of what browsing activities will be performed.
• Terminating other applications running on the device which may be also using the HTTP protocol.
• Opening the browser and clearing all the cookies and search histories.
• Running Wireshark and starting to record the HTTP messages.
• Performing a set of actions as planned.
• Taking notes of what exact actions are performed if there are changes.
• Stopping the recording and saving the trace when the browsing actions are completed.
• Examining and observing the saved HTTP messages line by line and taking notes of what/who the third party sites are and what types of information are being shared to them.

We examined the HTTP headers in each file (as shown in Figure 1) resulting from the experiments in order to identify the types of shared information and the types of third party sites. Let us have a look at how we observe the HTTP conversation between the browser, first party and third party site in Table 2.

(a) The first author visits Mediterranean chicken pasta salad recipe page on Taste.com.au. The browser makes an HTTP request to Taste.com.au in order to retrieve the page content.
(b) Taste.com.au receives the HTTP request and returns the HTTP response which contains the page content (in HTML format) as well as Javascript code. The browser executes the code that it requires to fetch an image content from another site imrworldwide.com.
(c) The browser then sends another HTTP request to imrworldwide.com for the image content to display on recipe page. Within this request, Taste.com.au is seen to leak the user’s search keyword, Mediterranean chicken salad pasta to a third party site (imrworldwide.com) in the HTTP header.
(d) imrworldwide.com then sends an HTTP response (contains 1x1 pixel image) and sets cookies in the browser.
### HTTP Messages

<table>
<thead>
<tr>
<th></th>
<th>HTTP Messages</th>
</tr>
</thead>
</table>
| (a) | GET /recipes/25158/mediterranean+chicken+pasta+salad HTTP/1.1
|    | Host: www.taste.com.au
|    | User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10.6; rv:8.0.1) Gecko/20100101 Firefox/8.0.1
|    | Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
|    | Accept-Language: en-us,en;q=0.5
|    | Referer: http://www.taste.com.au
|    | Cookie: PHPSESSID=mersn2toc44pvidn2ph61t1lc3...

| (b) | HTTP/1.1 200 OK
| Date: Fri, 02 Dec 2011 00:38:07 GMT
| Server: Apache/2.0.52 (CentOS)
| Content-Type: text/html;charset=utf-8
| ...
| ...
| Line-based text data: text/html
| ...
| ...
| ...
| <script type="text/javascript">
| ...
| ...
| ...;img src="http://secure-au.imrworldwide.com/...

| (c) | GET /cgi-bin/m?...
| Host: secure-au.imrworldwide.com
| User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10.6; rv:8.0.1) Gecko/20100101 Firefox/8.0.1

| (d) | HTTP/1.1 200 OK
| Date: Fri, 02 Dec 2011 00:37:16 GMT
| Set-Cookie: V5=cookie1; expires=Mon, 29-Nov-2021 00:37:16 GMT; domain=.imrworldwide.com; path=/cgi-bin
| ...
| Compuverse GIF, Version: GIF89a
| Screen width: 1
| Screen height: 1
| ...

### Table 2: First party site connects to third party site while also sharing user’s searched words

#### 3. Findings

Visits to a few websites result in thousands of lines of HTTP messages captured by Wireshark, either on the Windows, Linux or Mac OS. Similar to the finding reported by Krishnamurthy et al. (2007), the ad-block extension on the browser helps to limit the number of third party site connections, but it does not eliminate them. We observed that not all the advertisements were removed from the visited pages while the ad-block was on. We summarise our findings by first reporting on the level of information propagation going beyond one level of the third party sites which was not covered in the previous literature, then on the types of third party site being contacted at each propagation level, while also pinpointing the types of information being transferred from the first party site. We found that, user’s identifiable and non-identifiable information were being leaked to the third party sites while the user was visiting the first party sites. Two types of third party site were identified in this study: advertisers or data aggregator, and SNSs.
3.1. First level traverse: From first party sites to third party sites

3.1.1. Third party sites who are advertisers or data aggregators

It is not surprising and it appears to be consistent with the literature that first party sites usually transferred user's search keywords to third party sites who are advertisers or data aggregators. The connection is necessary to fetch the advertisement contents from the advertiser servers to display on first party site pages. In this case, for example, paying a single visit to some non-SNS sites such as Taste.com.au (for a vanilla cupcake recipe), Lyrics007 (for a song lyric: Yesterday by the Beatles), and 9 News (for an article: Google gets personal with search results), resulted in at least 10 connections to third party sites while also transmitting the user's searches to those sites (Table 3).

<table>
<thead>
<tr>
<th>taste.com.au (Recipe: vanilla cupcake)</th>
<th>Lyrics007 (Song title: Yesterday)</th>
<th>9 News (Article: Google gets personal with search results)</th>
</tr>
</thead>
<tbody>
<tr>
<td>News-static.com</td>
<td>Clickfuse.com</td>
<td>Msnportal.112.2o7.net</td>
</tr>
<tr>
<td>sops.news.com.au</td>
<td>addthis.com</td>
<td>imrworldwide.com</td>
</tr>
<tr>
<td>trakr-news.com.au</td>
<td>fastclick.net</td>
<td>m.adnxs.com</td>
</tr>
<tr>
<td>google-analytics.com</td>
<td>googlesyndication.com</td>
<td>bs.serving-sys.com</td>
</tr>
<tr>
<td>notebookmagazin.com</td>
<td>ringtonemaker.com</td>
<td>b.scorecardsearch.com</td>
</tr>
<tr>
<td>fashion.vogue.com.au</td>
<td>b.scorecardresearch.com</td>
<td>googlesyndication.com</td>
</tr>
<tr>
<td>doubleclick.net</td>
<td>apmefb.com</td>
<td>ringtonemaker.com</td>
</tr>
<tr>
<td>sunbeamfoods.com.au</td>
<td>rubiconproject.com</td>
<td>b.scorecardsearch.com</td>
</tr>
<tr>
<td>bs.serving-sys.com</td>
<td>rtbidder.net</td>
<td>ringtonemaker.com</td>
</tr>
<tr>
<td>jdn.monster.com</td>
<td>amazonaws.com</td>
<td>bs.serving-sys.com</td>
</tr>
<tr>
<td>facebook.com</td>
<td>jangonetwork.com</td>
<td>m.adnxs.com</td>
</tr>
<tr>
<td>twitter.com</td>
<td>advertising.com</td>
<td>b.scorecardsearch.com</td>
</tr>
<tr>
<td>getprice.com.au</td>
<td>abmr.net</td>
<td>googleapis.com</td>
</tr>
<tr>
<td>imrworldwide.com</td>
<td>googleapis.com</td>
<td>doubleclick.net</td>
</tr>
<tr>
<td>unica.com</td>
<td>api.google.com</td>
<td>facebook.com</td>
</tr>
<tr>
<td></td>
<td>twitter.com</td>
<td>twitter.com</td>
</tr>
</tbody>
</table>

Table 3: List of third party sites connected to by the browser while visiting first party sites

In terms of information leakages among SNSs (Table 4), there are two cases for Facebook. In the first case, connections were made only to Facebook CDN (Content Delivery Network) if no third party applications or advertisements on Facebook were clicked or used. In the second case, Facebook shared the user’s unique ID to Zynga (the online game company which hosts the Farmville application) when the user played Farmville from Facebook. Multiple requests started to be forwarded by Farmville to third party sites such as DoubleClick. However, transmission of the user’s Facebook ID from Farmville to advertisers was not observed. Meanwhile, although there were connections to third party sites, leaks or shares of identifiable information from Twitter were not observed, however Twitter page movements were. LinkedIn on the other hand, was seen to share or leak user’s information (in this case: LinkedIn ID and full name) to third party servers like Doubleclick (Table 5).
### 3.1.2. Third party sites who are SNSs

User's information and searches were seen to be transferred from first party sites to third party sites who are SNS. New SNS widgets, such as Facebook’s Like button, Twitter’s Tweet button and Google’s Plus One button, enabled site users to share content from other websites with their SNS friends (Facebook, 2012; Twitter, 2012; Google, 2012a). Facebook’s Like button also enabled site owners to have a view of the number of likes on their domain both daily and demographically. It was observed that when user visited sites embedded with those widgets, with or without logging-in into any SNS sites, those first party sites always sent HTTP requests to the SNS to populate the page with the SNS buttons. As shown in Table 3, Taste.com.au, Lyrics007, and 9 News, connected to at least 3 SNS: Facebook, Twitter and Google Plus because the SNS widgets reside on those first party sites. For example, in Table 6, Lyrics007 sent a request to Twitter for the widget while also sharing the user’s currently visited page. Twitter recognises the same visit from the cookies that were set when the user logged into her Twitter account.

<table>
<thead>
<tr>
<th>LinkedIn (No click on ads)</th>
<th>Twitter (No click on ads)</th>
<th>Facebook (Farmville application)</th>
<th>Facebook (No click on ads)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google-analytics.com</td>
<td>Google-analytics.com</td>
<td>Akamaihd.net</td>
<td>No third party sites</td>
</tr>
<tr>
<td>imrworldwide.com</td>
<td>twimg.com</td>
<td>googletagservices.com</td>
<td></td>
</tr>
<tr>
<td>quantserve.com</td>
<td></td>
<td>quantserve.com</td>
<td></td>
</tr>
<tr>
<td>b.scorecardresearch.com</td>
<td></td>
<td>doubleclick.net</td>
<td></td>
</tr>
<tr>
<td>doubleclick.net</td>
<td></td>
<td>rubiconproject.com</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>googleservices.com</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>quantserve.com</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>b.scorecardresearch.com</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>doubleclick.net</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>rtbidder.net</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>socialvi.be</td>
<td></td>
</tr>
</tbody>
</table>

### 3.2. Second level traverse: from third party to other third party sites

It was observed that the user’s information and the browser connection propagated beyond one level from the first party sites. The following illustrate this finding.
3.2.1 eBay case:
When the user visited eBay and searched or purchased an item, eBay shared the search keyword with other third party sites like Doubleclick. Doubleclick was also seen to share that search keyword with other third party sites, like amgdgt.comh and b.scorecardresearch.com (Table 7). User’s information was seen to be transmitted from eBay to third party site, and from that third party to another third party site as shown in Figure 2.

Table 7: The traverse of eBay user’s search keyword

![Figure 2: eBay traverse map](image-url)
3.2.2 Lyrics007 case:
Lyrics007 was similar to eBay in that there was a second level of information sharing to other third party sites. For example, the song title propagated from Lyrics007 to jangonetwork, then from jangonetwork to Doubleclick (Table 8). In addition, when user connected to Lyrics007(1), the browser connection went to rubiconproject.com(2), then from rubiconproject.com to w55c.net(3), and from w55c.net to bluekai.com (4)(Figure 3). However, information sharing ended in second level as shown in Figure 3.

<table>
<thead>
<tr>
<th>GET /00?... song title and artist name</th>
<th>GET /widgets.js HTTP/1.1</th>
<th>GET /gampad/ads?...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Host: jmn.jangonetwork.com</td>
<td>Host: partner.googleadservices.com</td>
<td>Host: pubads.g.doubleclick.net</td>
</tr>
<tr>
<td>User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10.6; rv:8.0.1) Gecko/20100101 Firefox/8.0.1</td>
<td>User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10.6; rv:8.0.1) Gecko/20100101 Firefox/8.0.1</td>
<td>User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10.6; rv:8.0.1) Gecko/20100101 Firefox/8.0.1</td>
</tr>
</tbody>
</table>

Table 8: Lyrics007 traverse map

Figure 3: Lyrics007 traverse map
Section 4: Discussion

First party sites were seen to leak user’s identifiable and non-identifiable information to third party sites. Third party sites in this study were identified to be advertisers or data aggregators and SNSs, particularly, Facebook, Twitter, and Google Plus (Table 9). Pieces of user’s information propagate from first party sites (both SNS and non-SNS) to third party sites, and also from third party sites to other third party sites. If the first party sites use or connect to the same third party sites, those third party sites are able to track user’s movements across different sites, among SNS and non-SNS, via the use of HTTP cookies. The tracking can be classified into two categories: the tracking by third party sites who are advertisers or data aggregators, and the tracking by third party sites who are SNSs.

<table>
<thead>
<tr>
<th>First party sites</th>
<th>The leaked information</th>
<th>Third party sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yahoo</td>
<td>User’s clickstream within Yahoo</td>
<td>Non-SNS</td>
</tr>
<tr>
<td>Gmail</td>
<td>Not observable</td>
<td>Not observable</td>
</tr>
<tr>
<td>Facebook (No ads or app clicked)</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Facebook (e.g. Farmville)</td>
<td>Facebook ID</td>
<td>Non-SNS</td>
</tr>
<tr>
<td>Twitter</td>
<td>User's searches and clickstream</td>
<td>Non-SNS</td>
</tr>
<tr>
<td>LinkedIn</td>
<td>User's clickstream, full name and ID</td>
<td>Non-SNS</td>
</tr>
<tr>
<td>Taste.com.au</td>
<td>User's searches and clickstream</td>
<td>SNS and non-SNS</td>
</tr>
<tr>
<td>Lyrics007</td>
<td>User's searches and clickstream</td>
<td>SNS and non-SNS</td>
</tr>
<tr>
<td>9 MSN News</td>
<td>User's searches and clickstream</td>
<td>SNS and non-SNS</td>
</tr>
<tr>
<td>Ebay</td>
<td>User's searches (e.g. hair-clips)</td>
<td>Non-SNS</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>YouTube</td>
<td>User's searches and clickstream</td>
<td>Google</td>
</tr>
<tr>
<td>Weatherzone</td>
<td>User's location and clickstream</td>
<td>SNS and non-SNS</td>
</tr>
</tbody>
</table>

Table 9: Summary of the information leakages in the study

Section 4.1: Tracking by advertisers or data aggregators as third party sites
Case: b.scorecardresearch
b.scorecardresearch is one of the most contacted third party sites among the visited sites in this case study. They are able to obtain both the user’s identifiable (in this case from LinkedIn: name and ID) and non-identifiable information (user’s searches). Therefore, within just one browsing session, b.scorecardresearch is able to associate a specific user with their online movements. In this case, for instance, b.scorecardresearch knows the first author’s name from LinkedIn and knows that she likes cooking (from her search for recipes) and listening to The Beatle.

Section 4.2: Tracking by SNS as third party sites
Through the use of the SNS widgets, it has been shown that the SNS are able to track users’ online movements not only within SNS but also across the different non-SNS that embed SNS widgets. SNSs themselves hold large amounts of personal information about a person’s life. There are two separate situations. First, without logging into any SNS during a browsing session, SNS can track a
user’s movements but not being able to identify a specific person. However, in the second situation, if user logged into any SNS account (Facebook, Twitter, or Google Plus, or even Gmail or Youtube account) while also browsing different websites, that SNS is able to combine the viewing habits with her profiles, within that browsing session.

**Case: Twitter**
Once a user logs into Twitter, Twitter sets cookies to remember the state (e.g. a user’s credential details). The cookies store information about a particular user (e.g. the user’s Twitter ID, name, and IP address). When user visits other sites that embed Tweet button, Twitter associates the visits with the same cookies value set while logging-in the account. Even after user logs out from her Twitter account, Twitter is still able to track her movement outside the SNS because the cookie values remain the same. From the perspective of Twitter, a Twitter user’s profile may also include his/her online browsing movements.

**Case: Facebook**
Facebook, on the other hand, is slightly different from Twitter in a sense that once user logs out from Facebook, the cookies associated with that person (which include a user’s name, e-mail address and ID) are destroyed. However, other cookies (e.g. the guest cookies) which are not directly associated with a particular person remain. Those cookies can still associate the browsed websites with the browsing devices (e.g. OS, browser, and IP address).

**Case: Google**
Google was observed to act differently in this study. The traffic associated with Gmail or Google Plus could not be observed in the HTTP messages. Rather, Google uses another protocol called TLS (Transport Layer Security) which is a cryptographic protocol providing secured communication over the internet. All the data are encrypted. However, we observed a form of tracking by Google via the cookie values used when the user later uses the Google search engine or YouTube. The majority of the first party sites within this study usually directly connect to Google or its other domains (e.g. google-syndication, or google-analytics) or its franchise (DoubleClick), so Google has the ability to track a user across different sites, both SNS and non-SNS, from the use of the same cookies.

Again, without logging into Gmail or Google Plus, there should not be any linkage of the user’s online search with a specific identifier. However, logging into its SNS or e-mail service will cause the user’s online activities (among first sites who use Plus one buttons or any sites resulted from Google search or any sites using Google ads) to be linkable with their Google profile via the HTTP cookie. We also noted that once user signed in into their Gmail account, they are automatically logged into Google Plus and YouTube. Though Gmail and Google Plus traffic data is not observable via the HTTP headers, the rest of the traffic via HTTP messages show that Google has so much details about a person’s life if that user logs into their Gmail account and stays logged-in during the browsing session.

**Conclusion**
Within this study, while non-SNSs share or leak non-identifiable information, they reveal a person’s browsing habits or searches to third party sites (e.g. advertisers and/or SNS). SNS like LinkedIn and Facebook (when the user clicks on advertisements or third party applications) share the user’s identifiable information (LinkedIn case: a user’s name and ID, Facebook case: Facebook’s ID) to third party sites. The traffic between the browser and Google Gmail and Google Plus services is not visible in the HTTP protocol because it is encrypted. However, based on the Google HTTP cookies, we observed that all the user's online movements can be tracked and linked to a specific user by Google if the user happens to log-in into any of its services (either Gmail, Google Plus, or YouTube) within the same browsing session.

It was also observed that both the user’s information (e.g. searches) and the browser connection propagate to more than just one level, traversing from first party sites to third party sites, and from those third party sites to other third party sites. There are two types of third party sites in this study: advertisers or data aggregators and SNSs such as Facebook, Twitter, and Google Plus. Both of them are able to track a user’s movement across different sites by the use of cookies and SNS widgets. With a piece of identifiable information from SNS (like in the case of LinkedIn: name and ID), the browsing habits or behaviours can be linked to a specific person.

Technologies that enable the tracking like HTTP cookies are not new. Users’ online movements as well as browsing behaviours could be tracked since cookies existed. It becomes problematic when a specific user can be identified. If a user never clears browser’s histories and cookies, the browsing profiles may consist of their online activities for 365 days of the year (some cookies can live up to 100 years). It appears that whenever a user visits any website (SNS and non-SNS), it is inevitable that they will leave their digital footprints within and across websites.

These results provide insight into the information leakages and the nature of the behavioural tracking within a browsing session where the first author of this article browses both SNS and non-SNS frequented by most people, and whose activities are common among most online users. It is important to look into the situation of information leakages from both SNS and non-SNS perspective, because the combination of the leaked information, both identifiable and non-identifiable (e.g. name and browsing activities), reveals so much about an individual. Although the findings of this study are not intended to generalise the facts about information gathering online, it does show that, within just one browsing session (as described in Section 2), the first author’s identifiable and non-identifiable information propagate to various third party sites who are advertisers and SNSs, and that these sites are able to track her browsing habits and combine with her identifiable information.

**Future work**

From the findings of this study, SNSs particularly Facebook, Twitter and Google Plus, and advertisers or data aggregators have the ability to track a user’s online movements across different websites. Online tracking is very common, however, SNSs are now holding not only large volumes of personal information provided by their online users, but also the information about their users’ movements outside SNS. There is not much on the moral aspects of SNS tracking; therefore, for our future work, we would like to analyse the privacy issues within the context of SNS tracking from the lens of a privacy framework like Contextual Integrity (CI).
CI was introduced by Helen Nissenbaum (Nissenbaum, 2004) as a framework for evaluating the impacts of a technology or system from a moral and political viewpoint. There are two norms in this privacy scheme known as: (a) norms of appropriateness (which dictate the types of information that is allowable to be divulged in a particular context), and (b) norms of distribution (which govern the flow of the information from one context to another). Contextual integrity of the information flow is maintained when both kinds of norms are respected; otherwise, a breach of privacy occurs. We would like to find out if the practice or ability of tracking users’ online movements within and outside SNSs by the SNS companies violate the norms of information flow, and thus users’ privacy.

References


About the authors

*Rath Kanha Sar* is a PhD candidate in the School of Computing and Mathematics at Charles Sturt University. Her research focuses on the privacy issues raised by the use of social network sites. She has given several presentations about her research including at international conferences.

Contact details: School of Computing and Mathematics, Charles Sturt University, Wagga Wagga, NSW 2678, Australia. E-mail: rsar@csu.edu.au

*Yeslam Al-Saggaf* is a Research Fellow at the Centre for Applied Philosophy and Public Ethics (CAPPE) and a Senior Lecturer in Information Technology at the School of Computing and Mathematics, Charles Sturt University. His research interests lie in the areas of privacy in social media and ICT ethics. He has published in those areas in a number of international refereed journals as well as presenting at a number of international conferences.

Contact details: School of Computing and Mathematics, Charles Sturt University, Wagga Wagga, NSW 2678, Australia. E-mail: yalsaggaf@csu.edu.au.