Private Information Leakage on the Mobile Web

Amanda Kirk, Stephen Rice, Zach Azar, Yipu Wang

amankirk@cs.du.edu, steprice@cs.du.edu, ZachAzar@cs.du.edu, yipu.wang@du.edu

11/18/13

COMP 4704: Foundations of Information Privacy

ABSTRACT

While leakage of private information to third-parties while browsing the web is well researched in academia [4] [5] [6], relatively little exploration has been made into the realm of privacy leakage on the mobile web. Building off of previous research by Krishnamurthy et al., [1] we explore the leakage of private information on non-online social networks (non-OSN's) via a mobile web browser. We select a subset of the categories from Krishnamurthy's previous, desktop-focused work and investigate five sites (per category) for provable private information leakage via a mobile device. Utilizing a variety of mobile devices and a PC running network analysis software, we crawl each site under the persona of a normal user. We perform standard actions such as creating accounts, navigating pages or performing searches, as well as site specific functions like sending dating messages on relationship web sites. Once we collect data for each crawl, we quantify the leakage we observe and compare our mobile-centered findings to known results on desktop web leakage. We find that our results for private information leakage on the mobile platform are comparable to Krishnamurthy's results from his previous study on traditional private information leakage via desktop web browsing for various non-OSN categories.

1. INTRODUCTION

The amount of private information that is being leaked to third-parties while browsing the web is a growing privacy concern amongst users and researchers alike [4] [5] [6]. Though previous research has confirmed and explored privacy leakage during desktop and mobile web browsing, third-parties continue to evolve their information gathering methods. Krishnamurthy et al., found that personally identifiable information is being leaked from an alarming percentage of traditional and mobile Online Social Networks [OSN's] [2] [3]. Furthermore, Krishnamurthy et al., found that while web browsing on a desktop, non-OSN's are also leaking personally identifiable information to various third-party networks [1]. Non-OSN's often encourage and allow users to create accounts so that they can interact with the site. Examples of non-OSN's are websites involving employment, photo sharing, travel, shopping, relationships, etc. With ever-growing advances in mobile technology and increasing access to the web from mobile devices, further research into the realm of mobile web tracking is necessary. We investigate leakage of private information to third-parties via HTTP requests across various non-OSN's. We explore what information is leaked, to whom it is being leaked, and in what manner it is being leaked to ensure a thorough analysis of mobile web leakage. We use Krishnamurthy's research on desktop non-OSN web leakage [1] to provide a base point of comparison for mobile privacy leakage.

2. BACKGROUND

There are multiple ways that information leakage can occur during normal web browsing (mobile or otherwise). This includes GET URLs, referer header field, cookies (including Flash cookies and growing cookies), transfer of sensitive information on unencrypted protocols, beacons, linking, use of JavaScript and Flash, and various other methods. For this study, we focus on leakage occurring in the GET URLs and referer header field for packets sent directly to third-parties. Many sites often use advertisements, pictures, or text provided from third-party websites. When your browser asks to retrieve web pages from first-parties, there are requests embedded in the returned HTML that encourage your browser to ask for ads, pictures, etc. directly from the third-party. This, by itself, does not constitute information leakage. However, personal information is often included in these requests and our browser unknowingly sends this sensitive information to the third-party directly. For example, a user is browsing amazon.com and the HTML includes an ad from doubleclick.net. The URL that Amazon provides to retrieve the ad using a GET method might have sensitive information encoded into it. Figure 1.1 is an example of this type of GET URL leakage where zip code, age, and potentially gender are leaked to the third-party via the GET URL. Another example of common leakage is using the referer header field. Often when we request content from a third-party, our browser is asked to include a referer header field that tells the third-party which first-party asked for the request. This is standard practice and does not leak private information by itself. However, sometimes the first-party will place private information in the referer header field so that it is sent to the third-party along with the request. Figure 1.2 is an example of this method where amazon.com has asked us to retrieve an advertisement from doubleclick.net again. In this example though, sensitive information is being leaked in the referer header field instead of the GET URL. Our study focuses on these two forms of leakage.

GET	http://ad.doubleclick.net/adj/...product;age=30;gnd=1;zip=80208...
Referer	http://www.amazon.com/...

Figure 1.1: Example of information leakage through the GET URL

GET	http://ad.doubleclick.net/?l=7654&sz=200x250...
Referer	http://www.amazon.com/hserver/age=30/zip=80208/gender=M/...

Figure 1.2: Example of information leakage through the referer header field

In a recent study by Krishnamurthy et al., [1], these forms of leakage were examined for non-OSN websites via desktop browsers. For various categories and subcategories from Alexa Top 500 (www.alexa.com), his team chose 10 sites based on popularity and usage. For each site, their team used various features of the site including creating an account, searching, editing their profile, etc. During their testing, they recorded all packets going in and out of the computer using a web proxy. Once they collected all of the data, they searched through the web traces looking for information leakage in the GET URL's and referer header fields. Table 1 is a summary of their published results. The second column in this table shows the number of sites who leaked information for each category (out of 10 sites). The remaining columns present how many sites leaked during those website actions for each category. For instance, 9 out of 10 health sites leaked information when the user performed a sensitive search. Krishnamurthy et al., performed a similar study for web leakage in OSN's on a desktop platform [3] as well as on a mobile platform [2]. Our study looks to mimic their study for non-OSN's on the mobile platform as their team has not yet conducted research in this area to our knowledge. Our aim is to select categories in a similar manner, navigate the sites using various mobile devices, record all packets during the site crawls, and examine the data looking for information leakage. We can then compare our results for mobile non-OSN web leakage to his results for desktop non-OSN web leakage to argue if there is more or less leakage when using a mobile browser for each category. We are also interested in seeing if there is more mobile specific leakage which takes advantage of browsing on a mobile device (GPS coordinates, access of contact lists, detection of other apps, etc.).

		Actions				
Category	Sites w/ Direct Leakage	Create Account	Account Login/ Navig.	View/ Edit Profile	Input Content	Sens. Search
Health	9	0	1	0	0	9
Travel	9	0	1	0	0	9
Employment	8	0	2	2	7	0
OSN	7	0	3	5	0	0
Arts	7	0	3	4	1	0
Relationships	7	0	3	2	2	0
News	5	0	5	0	0	0
PhotoShare	4	3	3	0	1	0
Sports	4	1	2	0	1	0
Shopping	3	0	2	0	2	0
AgeGroups	2	0	1	1	0	0
VideoGames	2	0	1	1	0	0
Tot. Sites/Cat.	67/12	4/2	27/12	15/6	14/6	18/2

Table 1: Leakage of Personal Information via Web Sites across Categories (Krishnamurthy, et. al [1])

3. METHODOLOGY

3.1 CATEGORIES AND SITES FOR STUDY

In order to build a solid and scientific foundation for our work, we begin our research by constructing data gathering and data analysis methodologies. We first determine what types of web sites we want to investigate for leakage and decide on what specific sites we actually use in our study. We begin with the list of Alexa Top 500 categories and subcategories that Krishnamurthy et al., used in previous work to choose what type of web sites we want to investigate for leakage. We customize this list for our study by narrowing it down based on the types of sites we expect to be accessed through a mobile browser and those who leaked information to third-parties on desktop browsers in previous studies. Our study focuses on four categories of non-OSN's: Travel, Shopping, Relationships, and Health. We then systemically choose five websites to examine for private information leakage for each of these categories. Our criterion for choosing these sites are as follows:

1. One of the top ranked sites on Alexa Top 500 [used popularly]
2. Does not require payment for registration
3. U.S. based, website in English
4. High possibility of being accessed by users on a mobile browser
5. Non-OSN (account requirement preferred)

The five sites we analyze for web leakage per category can be seen in Table 2.

Health	Relationships	Travel	Shopping
nih.gov	okcupid.com	agoda.com	amazon.com
webmd.com	pof.com	expedia.com	ebay.com
mayoclinic.com	kiss.com	booking.com	netflix.com
ncbi.nlm.nih.gov/pubmed	datehookup.com	hotels.com	walmart.com
myfitnesspal.com	friendfinder.com	tripadvisor.com	cvs.com

Table 2: Chosen Websites by Category

3.2 DATA GATHERING METHODOLOGY

To ensure consistency across our data set, we create a “test user” which will be used whenever personal information is required during our crawls. Our test persona has his own e-mail, user name, address, and other information that allow us to spot leakage more easily. By using this agreed upon set of private information, we not only ensure that our web crawl procedures are unified across categories but also ease the data analysis process. The initial steps for testing each site are to create an account and confirm the verification email (if needed). We keep all security settings at default. We then test the basic functionality of the website including page navigation, search and profile editing. There were also category specific actions, a few being:

- Shopping - shopping cart, purchase item, make a review, cancel order, track package, location of item
- Health- symptom query, symptom tree, read articles, search for doctors
- Travel- search itinerary, search cars/hotels/planes, select seats, location reviews
- Relationships-search for people, comment/poke/message, friend

We collected data for each website with Fiddler, a network analysis tool that enabled us to use our PC as a web proxy in order to observe HTTP requests being made between the mobile device and the internet. The PC’s were running either Windows 7 or Windows 8. We use three different major mobile operating systems in our study: iOS 7.0.2, Android 2.3.6 and Windows Phone 8 GDR2. Each crawl is done in isolation, starting with a blank web page and default browser settings. Every action made on a particular website is recorded in a “roadmap” detailing the type of action and its context, which allows us to backtrack and identify which stage of the web crawl we were involved in around the time of any one leak. Once each crawl is complete, the Fiddler session is saved for further analysis.

3.3 ANALYZING THE DATA

After gathering sufficient data, we investigate it for leakage of private information to third-parties. We search the captured HTTP requests/responses (GER URL's and referer header fields) looking for any form of leakage including name, zip code, GPS coordinates, and other private information. Since each crawl has hundreds if not thousands of packets to search through, we developed a way of searching the sessions for predetermined query strings. Using Fiddler's built-in "Find" feature, we can quickly search a set of packets for key words or phrases. We create a priority list of query strings to search for in each category of websites we included in our study. For example, we want to search all of the sessions in a web capture for the first name of our test user, to see if this information was leaked. This list is slightly different for each category of websites we included in our study since the information used during individual crawls varied due to the needs of the site. When we find leakage we note the manner of leakage, the corresponding action that we were doing on the website at the time of the leak using the aforementioned roadmap, what information was leaked, and the third-party that the information was leaked to. It should also be noted that due to the methodology of our search (manual) and the use of encryption by many third-parties, the leakage we find in our study represents a definitive lower bound on third-party leakage. Furthermore, as our study only examines direct leakage via HTTP requests (ignoring possible leakage via cookies and other methods), we suspect that there is far more leakage in our data than we have detected.

4. CONCLUSION

4.1 RESULTS

The following tables and graphs provide an overview of our collected results. Some graphs are broken down by category and will be accompanied by a single explanation. Table 4.1 reports how many sites (out of the total 5) leaked in each category. Similar to Krishnamurthy's results, the categories relationships, travel and health leaked in at least 80% of crawls, while Shopping only leaked in 40% of crawls.

Table 4.1: Leakage of Personal Information across Categories

Categories	Numbers of sites who leaked (out of 5)
Relationships	4
Travel	5
Shopping	2
Health	4

Tables 4.2 - 4.6 report all observed third parties as detected in each category. Both the relationship and health categories leaked to a significant number of third-parties, though it is notable that many of the third-parties are only used on one site. The two most prominent aggregators were google-analytics and doubleclick.

Table 4.2: Third-Party via Web Sites of Shopping

Third-Party	Number of sites
vi.ebaydesc.com	1
dev.virtualearth.net	1

Table 4.3: Third-Party via Web Sites of Travel

Third-Party	Number of sites
google-analytics.com	3
doubleclick.net	3
gstatic.com	2
facebook.com	2
criteo.com	2
(16 others)	1

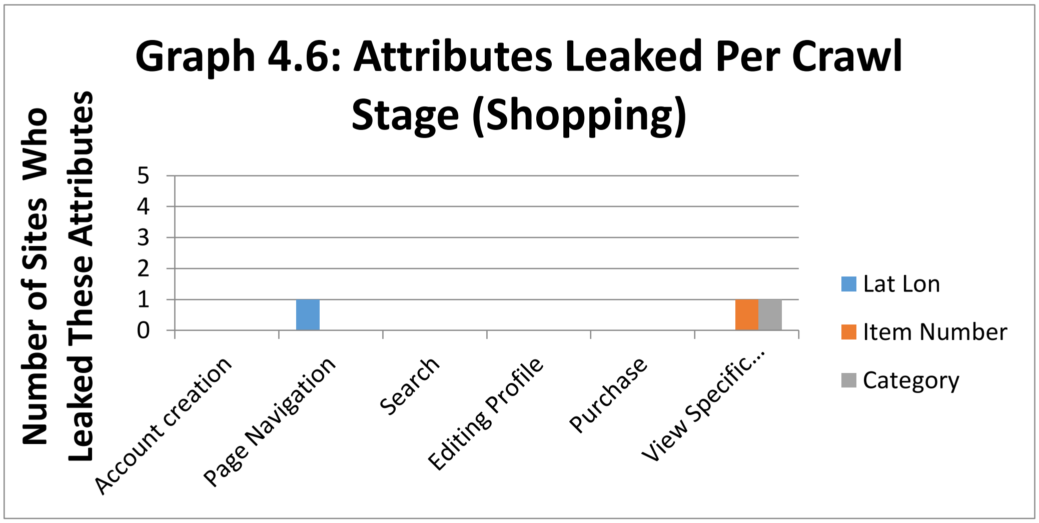
Table 4.5: Third-Party via Web Sites of Relationships

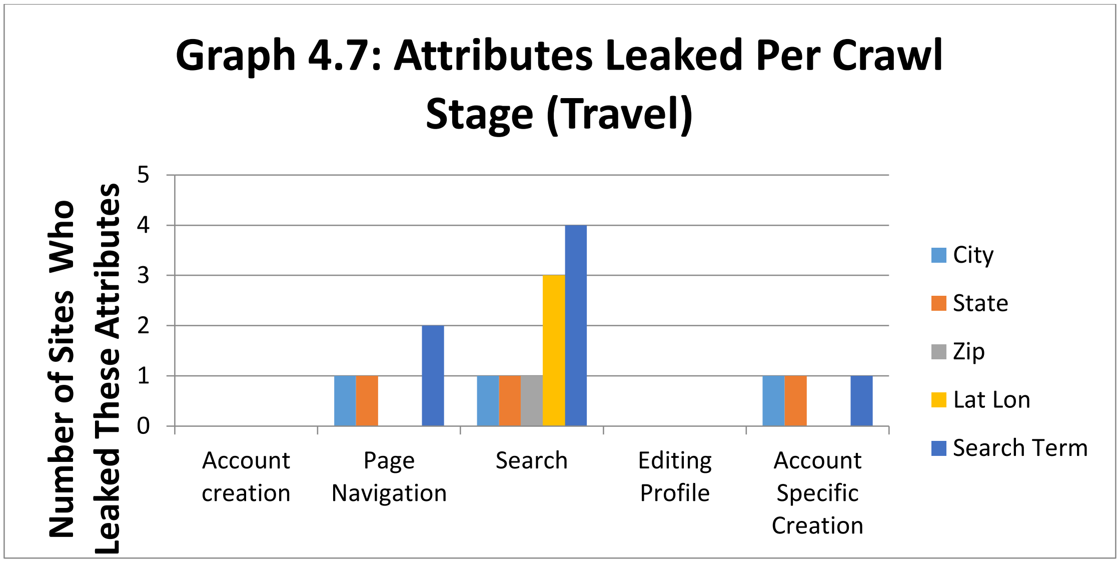
Third-Party	Number of sites
google-analytics.com	4
doubeclick.net	4
googlesyndication.com	3
(36 others)	1

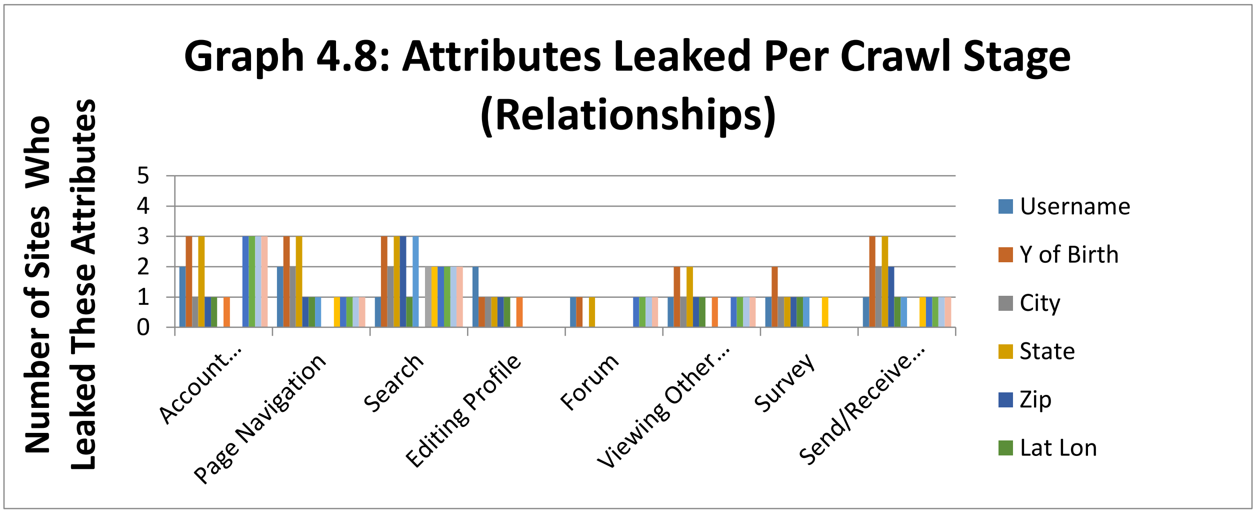
Table 4.6: Third-Party via Web Sites of Health

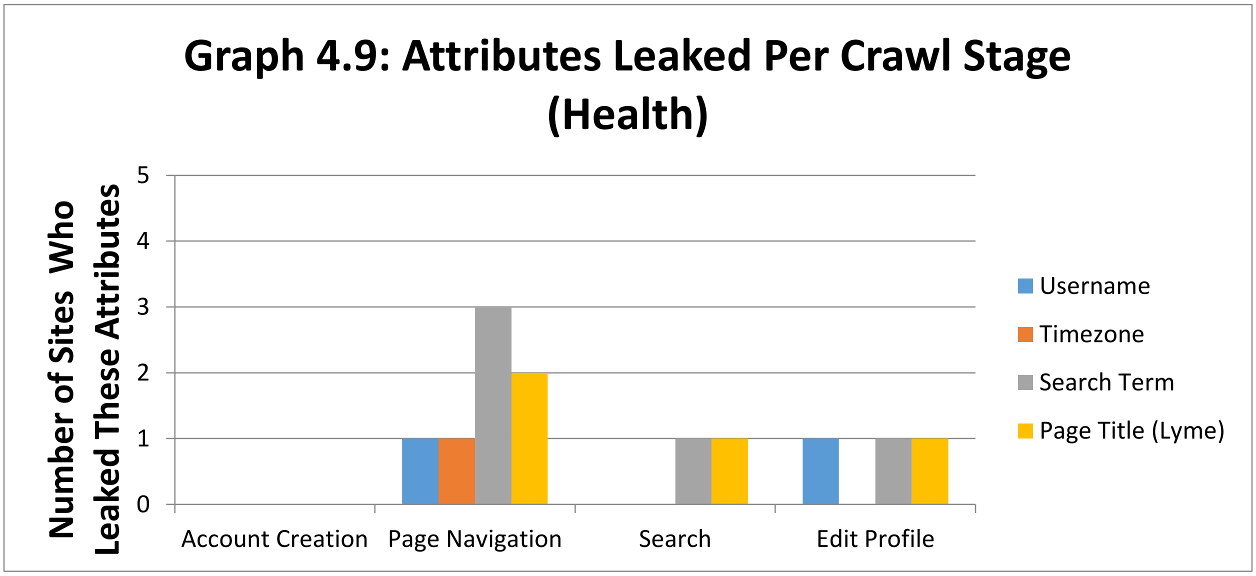
Third-Party	Number of sites
google-analytics.com	3
research.com	3
pointroll.com	2
adsafeprotected.com	2
newrelic.com	2
doubleclick.net	2
(40 others)	1

Graphs 4.6 - 4.9 report not only what attributes leaked but also what stage of the crawl it occurred in. As expected, the relationship category leaked more attributes during more crawl stages than any other category.


Graph 4.6: Attributes Leaked Per Crawl Stage (Shopping)


Graph 4.7: Attributes Leaked Per Crawl Stage (Travel)


Graph 4.8: Attributes Leaked Per Crawl Stage (Relationships)


Graph 4.9: Attributes Leaked Per Crawl Stage (Health)

Tables 4.10 - 4.13 report how often attributes leaked via the GET URL, the referer field, or with both methods. Note that the third column is a separate entity than the first two (i.e. an attribute was leaked in both the GET and Referrer in the same crawl). There does not seem to be a preferred manner of leakage for most third-party aggregators.

Table 4.10: Leak Method via Web Sites of Shopping

Attribute	GET URL	Referer Field	Both GET and Referer
Latitude & Longitude	0	1	0
Item Number	0	1	0
Category	0	1	0

Table 4.11: Leak Method via Web Sites of Travel

Attribute	GET URL	Referer Field	Both GET and Referer
City	0	0	3
State	0	0	3
Zip	0	1	0
Latitude & Longitude	0	1	2
Search Terms	0	0	5

Table 4.12: Leak Method via Web Sites of Health

Attribute	GET URL	Referer Field	Both GET and Referrer
Username	0	0	1
Time Zone	0	1	0
Search Terms	1	0	3
Page Title	1	0	3

Table 4.13: Leak Method via Web Sites of Relationships

Attribute	GET URL	Referer Field	Both GET and Referer
Username	1	0	1
Year of Birth	1	0	3
City	0	0	2
State	2	0	2
Zip	0	0	3
Latitude & Longitude	0	0	1
Gender	0	0	3
Email	1	0	0
Search Sex Seeking	0	0	2
Search Terms	0	0	2
Body Type	3	0	1
Education	3	0	1
Family	3	0	1
Occupation	3	0	1

4.2 RESULTS SUMMARY

Although not all categories leak the same amount, we observe some interesting trends for each category. Within the travel category, we see a surprising amount of leakage of exact GPS coordinates of the user in three of the five sites investigated. The travel category shows the search term as the most popular attribute being leaked to third-parties and searching to be the most common crawl stage where leakage occurs. The analysis of the relationship websites show that many attributes are leaked throughout multiple crawl stages. Only one of the relationship sites does not leak any information (and it is a premium site that only lets us navigate part of the site without a full subscription). The attributes that leak in the relationship category are very private (occupation, education, etc.). We observe that health sites fail to obscure information which can be classified as sensitive or non-sensitive depending on the study. For example, usernames and search terms are often leaked in page URL's. As these URL's are sent to third-parties (a standard measure), the information is considered to be "leaked". Leakage of these attributes could be malignant depending on the user's privacy preferences. The shopping category leaked very few sensitive attributes. We suspect that shopping websites tend to use better security when leaking information to third-parties. Large amounts of encrypted arguments were discovered to be sent to third-parties but we could not draw a firm conclusion. In Table 4.14, we see how our mobile leakage results compare to a previous study [1] on desktop web leakage on various non-OSN's. Although our study analyzes fewer sites than Krishnamurthy's original, our findings match his results within a 10% error rating. Therefore, our results for private information leakage on a mobile platform are comparable to results found in the realm of traditional privacy leakage.

Table 4.14: Percentage of Sites Who Leaked Per Category

	Krishnamurthy's Study: Desktop	Our Study: Mobile
Shopping	30%	40%
Health	90%	80%
Relationships	70%	80%
Travel	90%	100%

4.2 FUTURE WORK

Due to the surprisingly large variety of ways information can be leaked during mobile browsing, there are a number of aspects of our study that could be improved upon with more time and resources. First and foremost, our results would be much more strongly emphasized with a larger dataset to work with. This would require performing more web crawls and investigating more sites. As a more software oriented addition to this research, future work could be put into creating an automated program for the data gathering and analysis phases. Automating these steps in the research would result in more precise measurements, stronger results, and an overall improvement in the time it takes to perform the study. Another direction for further research would be to explore if the attributes being leaked change for accounts that are older (and thus have more private information associated with them). This could lead to a discovery of cross-pollination and linking across web sites for a single user. As mentioned earlier, our results also represent a lower bound on leakage due to the number of encrypted arguments that we are unable to decipher. The results of decrypting these arguments would allow future researchers to discover more (and perhaps more sensitive) forms of leakage. We did not often see information leaked alongside unique identifiers and we suspect that identifying information is being stored in cookies instead. Thus, exploring the use of HTTP Requests in conjunction with cookies and how the two forms of tracking utilize each other still needs to be explored on the mobile realm. Finally, a long-term study with the same methodology might uncover evidence that third-parties are linking and sharing information that they gather on specific users with one another.

REFERENCES:

[1] B. Krishnamurthy, K. Naryshkin and C.E. Wills, “Privacy leakage vs. Protection measures: the growing disconnect,” in *Proc. of the Web 2.0 Security and Privacy Workshop*., Oakland, 2011, pp. 1-10.

[2] B. Krishnamurthy and C.E. Wills, "Privacy leakage in mobile online social networks," in *Proc. of the Workshop on Online Social Networks*, Boston, MA, 2010, pp. 1-9.

[3] B. Krishnamurthy and C.E. Wills, "On the Leakage of Personally Identifiable Information via Online Social Networks," presented at the *Proc. of ACM SIGCOMM Workshop on Online Social Networks*, Barcelona, Spain, 2009.

[4] J. C. Mitchell and J. R. Mayer, "Third-Party Web Tracking: Policy and Technology," in *2012 IEEE Symposium on Security and Privacy*, San Francisco, CA, 2012, pp.413-427.

[5] B. Krishnamurthy and C.E. Wills, “Privacy diffusion on the web: a longitudinal perspective,” presented at the *Proc. of the 18th international conference on World wide web*, Madrid, Spain, 2009.

[6] J. Mayer, "Tracking the Trackers: where everybody knows your username," (CIS), [online] 2011, http://cyberlaw.stanford.edu/blog/2011/10/tracking-trackers-where-everybody-knows-your-username (Accessed: 9 October 2013).