On the Leakage of Personally Identifiable Information Via Online Social Networks

Balachander Krishnamurthy
AT&T Labs – Research
Florham Park, NJ USA
bala@research.att.com

Craig E. Wills
Worcester Polytechnic Institute
Worcester, MA USA
cew@cs.wpi.edu

Abstract

For purposes of this paper, we define "Personally identifiable information" (PII) as information which can be used to distinguish or trace an individual's identity either alone or when combined with other information that is linkable to a specific individual. The popularity of Online Social Networks (OSN) has accelerated the appearance of vast amounts of personal information on the Internet. Our research shows that it is possible for third-parties to link PII, which is leaked via OSNs, with user actions both within OSN sites and elsewhere on non-OSN sites. We refer to this ability to link PII and combine it with other information as "leakage". We have identified multiple ways by which such leakage occurs and discuss measures to prevent it.

Categories and Subject Descriptors

C.2 [**Computer-Communication Networks**]: Network Protocols—*applications*

General Terms

Measurement

Keywords

Online Social Networks, Privacy, Personally Identifiable Information

1. INTRODUCTION

For purposes of this paper, "Personally identifiable information" (PII) is defined as information which can be used to distinguish or trace an individual's identity either alone or when combined with other public information that is linkable to a specific individual. The growth in identity theft has increased concerns regarding unauthorized disclosure of PII. Over half a billion people are on various Online Social Networks (OSNs) and have made available a vast amount of personal information on these OSNs. OSN users make their information available (subject to the privacy policy of the OSN) to the authorized list of other OSN users, such as their 'friends'. Their profiles form a part of their online identity.

There has been a steady increase in the use of third-party servers, which provide content and advertisements for Web pages belonging to first-party servers. Some third-party servers are aggregators, which track and aggregate user viewing habits across different first-party servers, often via tracking cookies. Earlier, in [6] we showed that a few third-party tracking servers dominate across a number of popular Online Social Networks. Subsequently, in [7] we found that the penetration of the top-10 third-party servers across a large set of popular Web sites had grown from 40% in October 2005 to 70% in September 2008. A key question that has not been examined to our knowledge is whether PII belonging to any user is being leaked to these third-party servers via OSNs. Such leakage would imply that third-parties would not just know the viewing habits of *some* user, but would be able to associate these viewing habits with a specific person.

In this work we have found such leakage to occur and show how it happens via a combination of HTTP header information and cookies being sent to third-party aggregators. We show that *most users on OSNs are vulnerable to having their OSN identity information linked with tracking cookies.* [1] Unless an OSN user is aware of this leakage and has taken preventive measures, it is currently trivial to access the user's OSN page using the ID information. The two immediate consequences of such leakage: First, since tracking cookies have been gathered for several years from *non-OSN* sites as well, it is now possible for third-party aggregators to associate identity with those *past* accesses. Second, since users on OSNs will continue to visit OSN *and* non-OSN sites, such actions in the *future* are also liable to be linked with their OSN identity.

Tracking cookies are often opaque strings with hidden semantics known only to the party setting the cookie. As we also discovered, they may include visible identity information and if the same cookie is sent to an aggregator, it would constitute another vector of leakage. Due to the longer lifetime of tracking cookies, if the identity of the person is established even once, then aggregators could internally associate the cookie with the identity. As the same tracking cookie is sent from different Websites to the aggregator, the user's

WOSN'09, August 17, 2009, Barcelona, Spain.

[1] We have shared this information to all the OSNs we studied so that they may make informed decisions regarding preventative measures and subscriber notification.

movements around the Internet can now be tracked *not* just as an IP address, but be associated with the unique identifier used to store information about users on an OSN. This OSN identifier is a pointer to PII about the user.

Cookies and other tracking mechanisms on the Internet have been prevalent for a long time. The general claim of aggregators is that they create profiles of users based on their Internet behavior, but do not gather or record PII. Although we do not know that aggregators are recording PII, we demonstrate with this work that it is undeniable that information *is* available to them. Aggregators do not have to take any action to receive this information. As part of requests, they receive OSN identifiers with pointers to the PII or in some cases, directly receive pieces of PII. This PII information can be joined with information from tracking cookies obtained from the user's traversal to any site that triggers a visit to the same aggregator. The ability to link information across traversals on the Internet coupled with the wide range of daily actions performed by hundreds of millions of users on the Internet raises privacy issues, particularly to the extent users may not understand the consequences of having their PII information available to aggregators.

OSNs do have privacy policies on which OSN users rely when setting up and maintaining their account. These policies typically state that OSNs provide *non*-identifying information to third-parties as an aid in serving advertisements and other services. Many users, however may not understand the implications. The availability of a user's OSN identifier allows a third-party access to a user's name and other linkable PII that can identify a user. The goal of this work is not a legal examination of privacy policies, but to bring a technical examination of the observed leakage to the community's attention.

Section 2 enumerates pieces of personally identifiable information and examines the level of availability for these pieces across a number of OSNs. Section 3 describes our study of PII-related leakage in popular OSNs. Section 4 presents ways in which such leakage occurs across OSNs. Section 5 discusses techniques for possible protection against such leakage by the various parties involved in the transactions. We then look at preliminary work on the problem of PII leakage in non-OSN sites in Section 6. Section 7 concludes with a summary and description of future work.

2. AVAILABILITY OF PII IN OSNS

It is important to understand how the information provided to OSNs corresponds with PII and the nature of availability of such information to other users. PII is defined in [5] as referring to "information which can be used to distinguish or trace an individual's identity, such as their name, social security number, biometric records, etc. alone, or when combined with other personal or identifying information which is linked or linkable to a specific individual, such as date and place of birth, mother's maiden name, etc."

A recent report identifies a number of examples that may be considered PII [9] including: Name (full name, maiden name, mother's maiden name), personal identification number (e.g., Social Security Number), address (street or email address), telephone numbers, or personal characteristics (such as photographic images especially of face or other distinguishing characteristic, X-rays, fingerprints, or other biometric image). They can also include: asset information (IP or MAC address or a persistent static identifier that consistently links to a particular person or a small, well-defined group of people), or information identifying personally owned property (vehicle registration or identification number).

The report also includes examples of information about an individual that is linked or linkable to one of the above (e.g., date of birth, place of birth, race, religion, weight, activities, or employment, medical, education, or financial information). A well-known result in linking pieces of PII is that most Americans (87%) can be uniquely identified from a birth date, five-digit zip code, and gender [8]. A decade-old report [4] by the Federal Trade Commission in the U.S. specifically warned about the potential of linking profiles derived via tracking cookies and information about consumers obtained offline. It should be stressed that our work focuses on additional information obtained *online*.

With this understanding of PII, we analyze its availability and accessibility in the profile information for OSN users on popular OSNs. We used the 11 OSNs in our previous work [6]: Bebo, Digg, Facebook, Friendster, Hi5, Imeem, LiveJournal, MySpace, Orkut, Twitter and Xanga. We also included a 12th OSN for this study, LinkedIn, which is a popular professionals-oriented OSN.

The pieces of PII for an OSN user include: name (first and last), location (city), zip code, street address, email address, telephone numbers, and photos (both personal and as a set). We also include pieces of information about an individual that are linkable to one of the above: gender, birthday, age or birth year, schools, employer, friends and activities/interests. We only note availability if users are specifically asked for it as part of their OSN profile; otherwise we would not expect users to provide it. We do *not* process contents of OSN users' pages to see if they have additional personal information. Not all profile elements are filled in by users and entries may of course be false.

Table 1 shows the results of our analysis with the count of OSNs (out of 12) exhibiting the given degree of availability for each piece of PII (row). The first column indicates the number of OSNs where the piece of PII is available to all users of the OSN and the user *cannot* restrict access to it. This piece may also be available to non-users of the OSN—thus a primary source of concern. The second column shows the number of OSNs where the piece of PII is available to all users in the OSN via the default privacy settings, but the user can restrict access via these settings. The third column shows a count of OSNs where there is a piece of PII that users can fill out in their profile, but by default the value is not shown to everyone. The fourth column shows the count of OSNs where a piece of PII is not part of a user's profile and thus the information is not available unless the user goes out of their way to add it on their page.

The rows are sorted in decreasing order of availability and thus leakage (personal photos are available widely while street address are rarely present). The values in the first two columns raise more privacy concerns (hence the double vertical line). Prominence is given to them as we found in [6] that default privacy settings are generally permissive allowing access to strangers in all OSNs. We also found that despite privacy controls to limit access, between 55 and 90% of users in OSNs retain default settings for viewing of profile information and 80–97% for viewing of friends. The latter two columns suggests that some OSNs are concerned about the extent of private information that may be visible on

Table 1: PII Availability Counts in 12 OSNs

Piece of PII	Level of Availability			
	Always Available	Available by default	Unavailable by default	Always Unavailable
Personal Photo	9	2	1	0
Location	5	7	0	0
Gender	4	6	0	2
Name	5	6	1	0
Friends	1	10	1	0
Activities	2	8	0	2
Photo Set	0	9	0	3
Age/Birth Year	2	5	4	1
Schools	0	8	1	3
Employer	0	6	1	5
Birthday	0	4	7	1
Zip Code	0	0	10	2
Email Address	0	0	12	0
Phone Number	0	0	6	6
Street Address	0	0	4	8

OSNs. As we will see later, although some pieces of PII are unavailable to others in the OSN (the later rows) they may still leak via other means.

3. LEAKAGE STUDY METHODOLOGY

The concentration and default availability of pieces of PII for OSNs shown in Table 1 motivates our study to examine if and how PII is leaked via OSNs. We know that OSNs use a unique identifier for each of their users as a key for storing information about them. Such an identifier can also appear as part of a URI when user performs various actions on an OSN. For example, the identifier is often shown in the Request-URI when a user views or edits their OSN profile or clicks on a friend's picture. The use of this identifier is not a privacy concern if all interactions stay *within* the OSN, but as shown in [6] there is also interaction with third-party servers. If this interaction involves leakage of the unique identifier for a user then the third-party has a pointer to access PII of the user. The third-party may also have other information: tracking cookies with a long expiry period or source IP addresses, to join with the PII.

For the study, we log into each OSN and perform actions, such as accessing the user profile, that cause the OSN identifier to be displayed as part of the URI. We also click on displayed ads. While performing these actions we turn on the "Live HTTP Headers" [14] browser extension in Firefox, which displays HTTP request/response headers for all object retrievals. We analyze these headers to determine if any third-party servers are contacted, and if the user's OSN identifier or specific pieces of PII are visibly sent to the third-party servers via any HTTP header. Note that we will not detect if this information is sent via opaque strings.

A set of relevant request headers are shown in Figure 1 to illustrate an actual example of such a retrieval. Here `/pagead/test_domain.js` is retrieved from the server `googleads.g.doubleclick.net` as part of retrieving the set of objects needed to display content for a page on `myspace.com`.[2] As shown, the browser also includes the `Referer` (sic) header and a stored cookie belonging to `doubleclick.net`.

[2]In all examples, an OSN identifier of "123456789" or "jdoe" is substituted for the actual identifier in our study. Cookies and other strings are also anonymized.

```
GET /pagead/test_domain.js HTTP/1.1
Host: googleads.g.doubleclick.net
Referer: http://profile.myspace.com/index.cfm?
     fuseaction=user.viewprofile&friendid=123456789
Cookie:id=2015bdfb9ec||t=1234359834|et=730|cs=7aepmsks
```

Figure 1: Sample Leakage of OSN Identifier to a Third-Party

The `doubleclick.net` server is able to associate the user's identifier MySpace ("friendid" is the label used in URIs by MySpace to identify users, similar to 'id' or 'userid' used in other OSNs) with the DoubleClick cookie. Armed with this information the aggregator can join its "profile" of user accesses employing this cookie with any information available via the MySpace identifier.

4. LEAKAGE OF PII

Using the methodology described in Section 3 we examined the results of actions performed while logged onto each of the 12 OSNs in our study. We found four types of PII leakage involving the: 1) transmission of the OSN identifier to third-party servers from the OSN; 2) transmission of the OSN identifier to third-party servers via popular external applications 3) transmission of specific pieces of PII to third-party servers; and 4) linking of PII leakage within, across, and beyond OSNs. We now describe and show specific examples of how PII is transmitted to third-party aggregators.

4.1 Leakage of OSN Identifier

Our initial focus in the study is on the transmission of a user's OSN unique identifier to a third-party. Based on results in Table 1 the possession of this identifier allows a third-party to gain much PII information about a OSN user to join with the third-party profile information about a user's activity on non-OSN sites. Analyzing the request headers we obtain via the Live HTTP Headers extension, we find that the OSN identifier is transmitted to a third-party in at least three ways: the `Referer` header, the Request-URI, or a cookie. Examples for these three types of leakage are shown in Figure 2. Note that accesses to third-party servers are often triggered *without* explicit action (e.g., clicking on an advertisement) on the user's part.

```
GET /clk;203330889;26770264;z;u=ds&sv1=170988623...
Host: ad.doubleclick.net
Referer: http://www.facebook.com/profile.php?
     id=123456789&ref=name
Cookie: id=2015bdfb9ec||t=1234359834|et=730|cs=7aepmsks
```

(a) Via Referer Header

```
GET /__utm.gif?..utmhn=twitter.com&utmp=/profile/jdoe
Host: www.google-analytics.com
Referer: http://twitter.com/jdoe
```

(b) Via Request-URI

```
GET ...&g=http%3A//digg.com/users/jdoe&...
Host: z.digg.com
Referer: http://digg.com/users/jdoe
Cookie: s_sq=...http%25253A//digg.com/users/jdoe...
```

(c) Via Cookie

Figure 2: Leakage of OSN ID to a Third-Party

First, OSN identifiers can leak via the `Referer` header of a request when an identifier is part of the URI for a page.

OSNs typically include the identifier as part of a URI when showing the contents of any user's profile. As part of loading the contents for this page, the browser retrieves one or more objects from a third-party server. Each request contains the `Referer` header in the HTTP request, which passes along the OSN id. Figure 2a shows an example of this leakage where an object from the third-party `ad.doubleclick.net` is retrieved as part of a `www.facebook.com` page where the URI contains the OSN id and is thus included in the `Referer` header. In addition, the cookie for `doubleclick.net` is sent to the third-party server, which can now link this cookie with the OSN id. In testing, we observed similar examples of OSN id leakage to a third-party server via the `Referer` header in the presence of a third-party cookie for 9 of the 12 OSNs that we studied.

Second, OSN identifiers can leak to a third-party server via the request Request-URI. A typical example is shown in Figure 2b where a request to the analytics server `www.google-analytics.com` is made from a `twitter.com` page. This transmission not only allows the third-party to gather analytic information, but also to know the specific identifier of the user on the OSN. We observed such leakage for 5 of our 12 OSNs. The third-party domain `google-analytics.com` occurred in all five cases.

Third, OSN identifiers can leak to a third-party server via a first-party cookie when an OSN page contains objects from a server that appears to be part of the first-party domain, but actually belongs to a third-party aggregation server. We observed the increased use of such "hidden" third-party servers in [7] and observe similar use for OSNs in this work. In the example of Figure 2c, when we determine the authoritative DNS server for server `z.digg.com` we find that it is actually a server that is part of `omniture.com`, a large third-party tracking company [7]. Thus the browser includes the first-party cookie for `digg.com` in the request, which includes the OSN id, but the request is actually sent, because of the DNS mapping, to an `omniture.com` server. As the example shows, the OSN id is also sent via the Request-URI and `Referer` header, but this example is notable because it demonstrates another avenue of id leakage. We observed leakage of the OSN to such a "hidden" third-party server via a first-party cookie for 2 of the 12 OSNs.

In all, we observed the OSN id being leaked to a third-party server via one of these ways for 11 of the 12 OSNs. Such leakage allows the third-party to merge the OSN id with the profile of tracking information maintained by them.

The only OSN for which we did not observe such behavior is Orkut—part of the Google family of domains. Orkut requires a login via a Google account that is tracked via a Google cookie thus allowing the Orkut identifier to be directly associated with other `google.com` activity (e.g., search).

4.2 Leakage Via External Applications

External applications have become increasingly popular; Facebook alone has over 55,000 external applications. The applications are installed via the OSN but run on external servers not owned or operated by the OSNs themselves. The user is warned that downloading applications will result in the OSN sharing user-related information (including the identifier) with the external applications. Such sharing is required so that application providers can use them in API calls while interacting with the OSN. The user's social graph is accessible to the application only via the OSN and users interact with other OSN users (often their friends) via the application. Popular gaming applications and social interaction applications take advantage of the social graph to expand their reach quickly.

We observe that external applications of OSNs may themselves leak the OSN identifier to third-party aggregators. Once again, it is unclear if the OSN identifier needs to be made available to the external aggregator. While the leakage of the identifier in such cases is technically not the fault of the OSN, the user may not be aware of the secondary leakage occurring through external applications. Examples of leakage via requests involving external applications of MySpace and Facebook are shown in Figure 3.

```
GET /TLC/...
Host: view.atdmt.com
Referer: http://delb.opt.fimserve.com/
     adopt/...&puid=123456789&...
Cookie: AA002=123-456/789;...
```

(a) Via Referer Header

```
GET /...&utmhn=www.ilike.com&utmhid=1289997851&
     utmr=http://fb.ilike.com/facebook/
     auto_playlist_search?name=Bruce+Springsteen&...
     fb_sig_user=123456789&...
Host: www.google-analytics.com
Referer: http://www.ilike.com/player?
     app=fb&url=http://www.ilike.com/
     player/..._artistname/q=Bruce+Springsteen
```

(b) Via Request-URI

```
GET /track/?...&fb_sig_time=1236041837.3573&
     fb_sig_user=123456789&...
Host: adtracker.socialmedia.com
Referer: http://apps.facebook.com/kick_ass/...
Cookie: fbuserid=123456789;...=blog.socialmedia.com...
     cookname=anon; cookid=594...074; bbuserid=...;
```

(c) Via Request-URI and Cookie

Figure 3: Leakage of OSN ID to a Third-Party From an External Application

Figure 3a shows an example of a retrieved object from the third-party server `view.atdmt.com` with the MySpace identifier included in the `Referer` header. This retrieval follows a previous retrieval (not shown) where the use of a MySpace application causes the OSN identifier to be sent to the third-party server `delb.opt.fimserve.com` as part of the Request-URI. The example in Figure 3b shows a Facebook user's identifier being passed on by the popular external application "iLike" to a third-party aggregator `google-analytics.com` via the Request-URI. Figure 3c shows leakage via the Request-URI and `Cookie` header via a different application "Kickmania!" to an advertisement tracker `adtracker.socialmedia.com`.

4.3 Leakage of Pieces of PII

Beyond our initial focus on leakage of the OSN identifier, we also observe cases where pieces of PII are *directly* leaked to third-party servers via the Request-URI, `Referer` header and cookies. Figure 4 shows two such examples.

In Figure 4a, the third-party server `ads.sixapart.com` is directly given a user's age and gender via the Request-URI. This request is generated for an object on the user's profile page. This information is obtained from profile information stored for the user on `livejournal.com`. The third-party server also receives the OSN identifier via the

```
GET /show?gender=M&age=29&country=US&language=en...
Host: ads.sixapart.com
Referer: http://jdoe.livejournal.com/profile
```

(a) Age and Gender Via Request-URI

```
GET /st?ad_type=iframe&age=29&gender=M&e=&zip=11301&...
Host: ad.hi5.com
Referer: http://www.hi5.com/friend/profile/
     displaySameProfile.do?userid=123456789
Cookie: LoginInfo=M_AD_MI_MS|US_0_11301;
     Userid=123456789;Email=jdoe@email.com;
```

(b) Age, Gender, Zip and Email Via Request-URI and Cookie

Figure 4: Leakage of Pieces of PII to a Third-Party

`Referer` header, but obtains these two pieces of PII without even the need for a lookup.

In Figure 4b the server `ad.hi5.com` appears to be part of the `hi5.com` domain, but based on its authoritative DNS, it is actually served by the third-party domain `yieldmanager.com`. This third-party domain not only receives a user's age and gender, but also the user's zip code. These pieces of PII are supplied as part of the Request-URI. In addition, the first-party cookie for `hi5.com` contains the user's zip code and email address. Thus the third-party domain not only receives four pieces of PII, but the OSN is disclosing PII about the user that may not even be available to other users within the OSN. In our study, 2 of the 12 OSNs directly leak pieces of PII to third-parties via the Request-URI, `Referer` header and cookies.

4.4 Linking PII Leakage

Lastly, we examine possible linkages of PII leakage across, within, and beyond OSNs.

Across OSNs, once a third-party server is leaked PII information via one OSN, it may then also be leaked information via another OSN to which the same user belongs. For example, the cookie for `doubleclick.net` shown in the examples of Figure 1 and 2a means that DoubleClick can link the PII from across both MySpace and Facebook. This linkage is important because it not only allows the aggregator to mine PII from more than one OSN, but join this PII with the viewing behavior of this user.

Within an OSN, it is possible for a third-party server to not only obtain the OSN identifier for a user, but also the identifiers for the user's friends and other users of interest within the OSN. For example, a user viewing a friend's profile will leak that friend's OSN identifier.

Finally looking beyond the OSN, the use of a third-party tracking cookie allows the PII available from the OSN to be linked with other online user activity. For example, Figure 5 shows a retrieval from a site that a user may not want to be known to others, yet is linked to the same cookie as used to access MySpace and Facebook.

```
GET /pagead/ads?client=ca-primedia-premium_js&...
Host: googleads.g.doubleclick.net
Referer: http://pregnancy.about.com/
Cookie: id=2015bdfb9ec||t=1234359834|et=730|cs=7aepmsks
```

Figure 5: Example of Third-Party Cookie for Non-OSN Server

5. PROTECTION AGAINST PII LEAKAGE

We have demonstrated a variety of scenarios whereby OSN identifiers and PII present on the corresponding user profiles leak via different OSNs. We now examine the parties involved in the leakage and the ways by which they can help prevent it. There are primarily four parties involved in the series of transactions: the user, third-party aggregators, the OSN, and any external applications accessed via the OSN.

Users ability to block leakage of PII range from the draconian, albeit effective, one of not disclosing any in the first place to being highly selective about the type and nature of personal information shared. Facebook applications have been created to increase awareness of information that could be used in security questions [10] and provide mechanisms for additional privacy protection [11]. Known privacy protection techniques at the browser include filtering out HTTP headers (e.g., `Referer`, `Cookie`), and refusing third-party cookies. The potential problem with the `Referer` header to leak private information was identified in 1996 (!) in the HTTP/1.0 specification [2]:

> Because the source of a link may be private information or may reveal an otherwise private information source, it is strongly recommended that the user be able to select whether or not the Referer field is sent.

Firefox allows direct blocking of `Referer` header [3] or as add-on with more per-site control [1]. With user customization, some actions may cause further accesses to be affected. For example, some servers check the `Referer` header before they answer any requests, in an attempt to prevent their content from being linked to or embedded elsewhere. Protection techniques could be deployed at a proxy [12, 13] to benefit all users behind it. Recently, the HTTP Working Group has had discussions on new headers (such as the `Origin` header) to replace the `Referer` header.[3] Only the information needed for identifying the principal that initiated the request would be included and path or query portions of the Request-URI are excluded. The proposal has not advanced significantly. Additionally, as we have demonstrated, even if the user filtered the `Referer` header and blocked cookies, the OSN identifier is also leaked in the `GET` or `POST` request via the Request-URI.

Second, aggregators could filter out any PII-related headers that arrive at their servers and ensure that tracking mechanisms are clean of PII at all times. Publishing the hidden semantics of cookies could work as a confidence building measure; the current opaque string model implies that users will not know if different cookies received (e.g., after deleting older cookies) are being correlated.

Third, OSNs could ensure that a wide range of privacy measures are available to members. Providing strong privacy protection by default allows an OSN to distinguish itself from other competing OSNs. Techniques at OSNs are in reality much easier. Most leakage identified in this study originated from the OSN allowing the internal user identifier to be visible to the browser unnecessarily leading to the population of the `Referer` header. A straightforward solution is to strip any visible URI of userid information. Alternately the OSN could keep a session-specific value for the user's identifier or maintain an internal hash table of the ID and present a dynamically generated opaque string to

[3]Currently available at `http://tools.ietf.org/html/draft-abarth-origin-00`.

the browser. If the opaque string is included in the `Referer` header by the browser, no information is leaked as the external site will not be able to use the opaque string to associate with the user and thus their PII.

In some cases Facebook inserts a '`#`' character before the id field in its Request-URI. Since some browsers only retain the portion before '`#`' in a URI to be used in `Referer` headers and such, this may reduce chances of leakage. However, as our examples have shown, Facebook does not consistently follow this technique; even when consistently followed, other (non-`Referer` header related) leakage mechanisms outlined will continue to occur.

The fourth party, external applications, allow the OSN identifier to be passed through to external aggregators. They could use one of the methods outlined above to strip the id or remap it internally.

6. LEAKAGE VIA NON-OSN SITES

Although we focus on OSNs in this study, it should be obvious that the manner of leakage could affect users who have accounts and PII on other sites. Sites related to e-commerce, travel, and news services, maintain information about registered users. Some of these sites do use transient session-specific identifiers, which are less prone to identifying an individual compared with persistent identifiers of OSNs. Yet, the sites may embed pieces of PII such as email addresses and location within cookies or Request-URIs.

We have carried out a *preliminary* examination of several popular commercial sites for which we have readily available access. These include books, newspaper, travel, micropayment, and e-commerce sites. We identified a news site that leaks user email addresses to at least three separate third-party aggregators. A travel site embeds a user's first name and default airport in its cookies, which is therefore leaked to any third-party server hiding within the domain name of the travel site. By and large we did *not* observe leakage of user's login identifier via the `Referer` header, the Cookie, or the Request-URI. It should be noted that even if the user's identifier had leaked, the associated profile information about the user will not be available to the aggregator without the corresponding password.

Our preliminary examination should not be taken as the final answer on this issue. A thorough understanding of the scope of the problem along with steps for preventing leakage in general remains a primary concern. Any protection technique must effectively ensure de-identification between a user's identity prior to any external communication on any site that requires logging in—OSN or otherwise.

7. CONCLUSION

The results of our study clearly show that the indirect leakage of PII via OSN identifiers to third-party aggregation servers is happening. OSNs in our study consistently demonstrate leakage of user identifier information to one or more third-parties via Request-URIs, `Referer` headers and cookies. In addition, two of the OSNs directly leak pieces of PII to third parties with one of the OSNs leaking zip code and email information about users that may not be even publicly available within the OSN itself. We also observe that this leakage extends to external OSN applications, which not only have access to user profile information, but leak a user's OSN identifier to other third parties. It should be noted that there may be private contractual agreements between aggregators and OSNs that forbid aggregators from using any information they may receive as a result of user's interaction with an OSN.

OSNs are in the best position to prevent such leakage by eliminating OSN identifiers from the Request-URI and consequently the `Referer` header. This elimination can be done directly or by mapping an OSN identifier to a session-specific value. Users have some means for limiting PII leakage via what information they provide to the OSN or browser/proxy techniques to control use of the `Referer` header and cookies. However, these controls may break accesses to other sites or not completely eliminate PII leakage via OSNs.

A clear direction for future work is to understand the bigger picture of PII leakage to third parties. We have performed a preliminary examination of PII leakage for non-OSN sites and found a couple of instances where pieces of PII were leaked to third-parties. We plan to undertake a more extensive examination of this issue along with steps that can be taken to prevent leakage of private information.

Acknowledgments

We would like to thank Steven Bellovin, Graham Cormode, Jeff Mogul, Raj Savoor, Josh Elman, and the anonymous reviewers for their comments.

8. REFERENCES

[1] James Abbatiello. Refcontrol. Firefox Add-on. `https://addons.mozilla.org/en-US/firefox/addon/953`.

[2] T. Berners-Lee, R. Fielding, and H. Frystyk. Hypertext Transfer Protocol — HTTP/1.0. RFC 1945, IETF, May 1996. Defines current usage of HTTP/1.0. `http://www.rfc-editor.org/rfc/rfc1945.txt`.

[3] The cafes: Privacy tip #3: Block referer headers in Firefox, October 2006. `http://cafe.elharo.com/privacy/privacy-tip-3-block-referer-headers-in-firefox/`.

[4] Online profiling: A report to congress, July 2000. Federal Trade Commission. `http://www.ftc.gov/os/2000/07/onlineprofiling.htm`.

[5] Clay Johnson III. Safeguarding against and responding to the breach of personally identifiable information, May 22 2007. Office of Management and Budget Memorandum. `http://www.whitehouse.gov/omb/memoranda/fy2007/m07-16.pdf`.

[6] Balachander Krishnamurthy and Craig E. Wills. Characterizing privacy in online social networks. In *Proceedings of the Workshop on Online Social Networks*, pages 37–42, Seattle, WA USA, August 2008. ACM.

[7] Balachander Krishnamurthy and Craig E. Wills. Privacy diffusion on the web: A longitudinal perspective. In *Procs World Wide Web Conference, Madrid, Spain*, April 2009. `http://www.research.att.com/~bala/papers/www09.pdf`.

[8] Bradley Malin. Betrayed by my shadow: Learning data identify via trail matching. *Journal of Privacy Technology*, June 2005.

[9] Erika McCallister, Tim Grance, and Karen Scanfone. Guide to protecting the confidentiality of personally identifiable information (PII) (draft), January 2009. NIST Special Publication 800-122. `http://csrc.nist.gov/publications/drafts/800-122/Draft-SP800-122.pdf`.

[10] Privacy guard. Facebook Application. `http://apps.facebook.com/privacyguard/`.

[11] Privacy protector. Facebook Application. `http://apps.facebook.com/privacyprotector/`.

[12] Privoxy. `http://www.privoxy.org/`.

[13] Proxify anonymous proxy. `http://proxify.com/`.

[14] Daniel Savard. LiveHTTPHeaders. Firefox Add-on. `http://livehttpheaders.mozdev.org/`.