On the Leakage of Personally Identifiable Information Via Online Social Networks

Balachander Krishnamurthy
AT&T Labs – Research

Craig E. Wills
Worcester Polytechnic Institute

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

[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 the browser. If the opaque string is included in the `Referer`

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

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/`.

Privacy Diffusion on the Web: A Longitudinal Perspective

Balachander Krishnamurthy
AT&T Labs–Research

Craig E. Wills
Worcester Polytechnic Institute

ABSTRACT

For the last few years we have studied the diff sion of private information about sers as they visit various Web sites triggering data gathering aggregation by third parties. This paper reports on our longitudinal study consisting of multiple snapshots of our examination of s ch diffusion over fo r years. We examine the various technical ways by which third-party aggregators acquire data and the depth of user-related information acquired. We study techniques for protecting against this privacy diff sion as well as limitations of such techniques. We introduce the concept of secondary privacy damage.

O r res lts show increasing aggregation of user-related data by a steadily decreasing number of entities. A handful of companies are able to track sers' movement across almost all of the popular Web sites. Virtually all the protection techniques have significant limitations highlighting the seriousness of the problem and the need for alternate solutions.

Categories and Subject Descriptors

C.2 [Computer-Comm nication Networks]: Network Protocols—applications

General Terms

Measurement

Keywords

Privacy, Privacy Enhancing Technologies

1. INTRODUCTION

The European Union's Privacy directive [7] defines an "identifiable person" as "one who can be identified, directly or indirectly, by reference to an identification n mber or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity." It is well known that combinations of different data elements can lead to niquely identifying a person. Privacy literature has introd ced terms like de-identification (stripping identity information from data) and re-identification (ability to relate supposedly anonymous data with actual identities). Concerns abo t user privacy have risen dramatically with the increased dependence on the Internet for a wide variety of daily transactions that leave trails to be left in many locations. Web terms like cookies are widely known and modern browsers provide privacy protection choices.

A common refrain is that any perceived loss of privacy emanating from normal actions on the Internet does not amount to actual loss of privacy as 'personally identifiable information' (PII) is not gathered, assembled, or retained. While evidence for this claim is not available neither is there convincing proof that the data that has been gathered over the 16 years of Web's existence amounts to PII. The widespread popularity of the Web indicates that most users either do not know or do not care about any perceived loss of privacy. However, recent concerns about identity theft and news stories of privacy breaches are increasingly changing how sers think abo t their privacy.

Our thesis is that there are causes for concerns about potential loss of PII based on the growth and aggregation of information tracking res lting from sers' activities on the Web. Gathering a certain amount of private information is essential for applications: it is impossible to sell books over the Internet without obtaining name, credit card information, and address. S ch e-commerce sites are often diligent with the s pplied information for practical and legal reasons. However, significant amount of in-depth tracking by a large fraction of popular (and not so popular) Web sites is also widespread.

We do not know if the data that has been and is being gathered can definitely be translated to PII; however it is hard to ignore the concentration and breadth of data being acquired. Aggregation of data by sophisticated technical means has been augmented recently by direct acquisitions of companies (along with their longitudinal data).

We do not claim that all data acquisition is of concern, nor do we assert that sers should block private information from being gathered in all cases. It is important for sers to know what is being gathered, how, and whether it is necessary. Ideally, users should reach a mod s vivendi whereby they consent to what is being tracked by selected sites to an approved extent. If it is possible for them to interact without their privacy being diff sed they should be able to do so. O r work is an initial step in trying to move towards s ch an informed consens s that balances the needs of sites and the legitimate privacy needs of sers.

Yet, there is little data about s ch privacy diffusion on the Internet resulting from individ al ser's actions involving visits to popular Web sites. In earlier work, we took a first c t at examining the problem of privacy diffusion on

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

the Web [13, 11]. In this paper, we present a longitudinal perspective of o r study spanning four years exploring the nature and extent of tracking of ser-related information by a large set of popular Web sites. O rs is the first s ch study to examine privacy diff sion over time that covers a broad set of technologies sed for tracking and the potential for various meas res against s ch tracking.

The organization for the paper is as follows. Section 2 en - merates the list of privacy related data elements currently being tracked on the Web and the techniques sed for such tracking. Section 3 describes the methodology of o r longitudinal study together with its technical scope. Section 4 presents the res lts of our longitudinal study and reasonable inferences that can be drawn. Section 5 demonstrates limitations of current privacy protection techniques. Section 6 presents arguments of how PII could be gleaned by combining the data elements already being gathered with ambient information and other popular applications that are not covered in our study. Section 7 raises a new issue of secondary privacy damage where the actions of one user can leak information about another user an aggregator of information. We concl de in Section 8 with a s mmary and a look at future work. We note that the code we used to gather data is available for repeating o r experiments on any s bset of Web sites of interest to readers.

2. PRIVACY ELEMENTS

We now en merate the list of privacy related data elements c rrently being tracked on the Web and the techniques sed for such tracking. While our list is not exhaustive, we capture the most common elements and techniques.

A user's visit to a single Web site (what we term a first-party site) often res lts in multiple HTTP requests being sent to n merous servers under the control of different administrative entities. Some requests are necessary to obtain the content being requested from the site owner's servers or Content Distrib tion Network (CDN) sites, while others are needed to fetch advertisements. Yet others are p rely for the purpose of tracking a user's movements on the Web. All sites visited other than the first party are termed as third-party sites. Although CDNs are indeed capable of tracking ser's movements, we discount their role when they distribute content on behalf of the first parties. We also note that some tracking is seful: cookies allow users to visit the site again and have their profile re-used to avoid having to re-enter information. Note that both first and third parties send cookies. Other tracking mechanisms are j stified by the claim that they enhance the user's experience; e.g., the use of JavaScript.

Behavioral tracking is one of the oldest techniques employed on the Web. Behavioral tracking allows for monitoring user Web accesses across m ltiple nrelated Web sites. A common application is to see if a partic lar ad displayed on a site res lted in the user clicking on it. The common technique is to se a cookie that can be correlated across multiple sites; the aggregator knows that it is the same user who has visited these sites. The definition of a ' ser' is somewhat nebulo s: it could be simply the IP address present in the client HTTP request. But in combination with simple ambient information it may be possible to ensure that it represents a single ser rather than multiple people sending requests from that IP address. For example, examining the access patterns over time, and the time periods and frequency of accesses, it may be easy to distinguish sers even if multiple sers are behind the same address. Web bugs (the 1x1 pixel GIF images) are another way to extract information abo t sites sers are visiting. The advantage of behavioral tracking is th s the ability to create a profile of a ser [16]. Use of tracking cookies is fairly ubiquitous [19] and there are known techniques to avoid them [22].

Some third parties provide Web analytics services for traffic measurement, user characterization, connectivity and geolocation services. Often a JavaScript file is downloaded to a client browser which in addition to the computation creates and updates first-party cookies. The scripts send information back to the third-party site through identifying URLs (containing characters like '?', '=', or '&') that are sed to pass parameter val es and information to the server. Note that JavaScript does not have to be downloaded as a separate object but can be present inline in the original HTML downloaded.

Cookies, being opaque strings can encode any information that a sending server desires and can change over time. JavaScript, being exec table code, can carry out computations at the client's side although it has limited access to ser data. Scripts do have access to information in the browser incl ding cached objects and the history of visited links [10]. Along with cookies and res lts of JavaScript exec tion, the tracking sites have all the regular information available in a typical HTTP request: sender's IP address, ser-agent software information, c rrent and previous URL (via Referer header), email address (From header), lang age preference (Accept-Lang age header), etc. Beyond these, depending on the site visited search strings, passwords, account n mbers, etc. may also be available, although typically only to the first-party site.

Behavioral tracking sites like doubleclick.net and tacoda.net have been around for well over a decade (although both have been recently acquired by larger companies). Prominent Web analytics domains are google-analytics.com, quantserve.com and omniture.com.

3. LONGITUDINAL STUDY

In the following we describe the methodology of o r longitudinal study along with its technical scope. Our study involved downloading aro nd 1200 popular Web sites (from more than 1000 nique servers) over five epochs of time between October 2005 and September 2008 and examining the additional Web sites visited by the browser. The study was automated using a Firefox extension [6] to drive the retrieval of the each first-party site while the extension recorded all of the resulting third-party sites visited[1]. We also examined the presence of cookies, JavaScript, and identifying URLs in the downloaded pages. The study set incl ded English-lang age sites obtained across various categories from Alexa's popular sites [3], first sed in [12]. Our study used the same data set of pop lar Web sites over all epochs, although we also examined the impact of using the c rrent Web site membership for the Alexa categories.

We also examine two important s bsets of the broadly popular Web sites: a) consumer sites, where users do not just browse but supply additional personal information s ch as credit card n mbers and b) fiduciary sites, where sers

[1]A proxy was used to record visited sites in the October 2005 epoch.

provide a variety of personal information incl ding bank acco nt numbers, and other personally identifiable information.

In analyzing the use of third-party sites across this set of first-party sites, which are identified based on their server, we refined the approach defined in [13] to merge third-party servers from the same organization. In that work, we sed a "domain" approach where third-party servers with the same 2nd-level domain are merged into a single third-party domain[2]. Thus the third-party servers walmartcom.112.2o7.net and timecom.122.2o7.net are merged into the 2o7.net third-party domain.

The weakness of this approach is that it fails to capture cases where what appeared to be a server in one organization (e.g. w88.go.com) was actually a DNS CNAME alias to a server (go.com.112.2o7.net) in another organization (Omniture). We found these type of relationships could be captured with an "adns" approach where all third-party servers sharing the same set of authoritative DNS servers (ADNSs) were merged into the same third-party.

In this work, we found neither of this approaches alone to be satisfactory for merging third-party servers together for analysis. While the ADNS approach overcomes weaknesses in the domain approach it has other drawbacks. For example, DNS for some third-party servers is provided by DNS services, s ch as ltradns.net. These services do not represent the source of the content. Similar iss es arise with content distributed networks (CDNs), which were originally developed to deliver content behalf of first-party servers. Increasingly CDNs are being sed to serve content, s ch as JavaScript or images with cookies attached, on behalf of other third-party servers. For example, an Akamai server is sed to serve content for the third-party server pixel.quantserve.com. This third-party content belongs to quantserve.com and should not be gro ped with all other content of servers with an Akamai ADNS.

Because of these shortcomings we use a refined approach in this work, which we call the "root" domain, to gro p servers. We start with the domain of the third-party server, but we also obtain the ADNS of the third-party server as well as the ADNS of the first-party server. If the ADNS of the third-party server is not the same as that of the first-party server and the ADNS is not that of a known CDN or DNS service then we se the ADNS as the root domain. Th s the root domain of www.google-analytics.com is google-analytics.com and the root domain of pixel.quantserve.com is q antserve.com even though its content is served by the Akamai CDN. Similarly the root domain of adopt.specificclick.net is specificclick.net as its ADNS is from the ultradns.net domain. Finally, the root domain of w88.go.com is omniture.com beca se its content is served by an Omniture server.

We se these third-party root domains to examine the diffusion of information about user viewing habits across o r set of popular first-party sites. In [13], we defined the notion of a privacy footprint to examine this diff sion. The footprint metric shows the n mber and diversity of third-party sites visited as a result of visiting first-party sites. Here, we track this footprint longitudinally by examining the penetration of the most used third-party root domains, which are in a position to aggregate information abo t user viewing habits, across the set of first-party sites. We also examine the depth of third-party tracking in terms of the n mber of these third-party domains that are present on each first-party site. Finally, we show the impact of a new factor: economic acquisition, where one aggregator p rchases another—instantly and sometimes significantly increasing its footprint.

[2]In cases where the Top-Level Domain (TLD) is a country code and the TLD is s bdivided using recognizable domains such as "com" or "co" then the domain approach groups servers according to the 3rd-level domain.

4. RESULTS

This section describes res lts from sing the basic methodology for data gathering and analysis described in the previo s section.

4.1 Longitudinal Results of Top Third-Party Root Domains

Focusing on the penetration of third-party root domains amongst the set of first-party servers in o r basic test data set, Figure 1 first shows the cumulative penetration of the top-10 root domains at the time of each of the five epochs in o r longitudinal study. The res lts show that the top-10 domains were sed by 40% of first-party servers in Oct'05, b t had extended to 70% of the first-party servers by Sep'08.

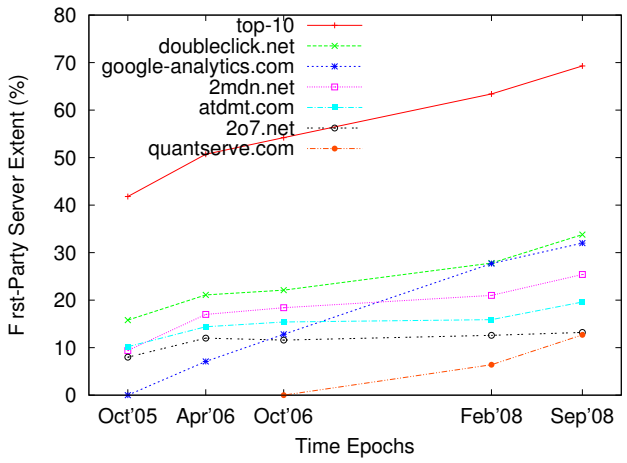

Fig re 1: Extent of Top-10 Third-Party Root Domains in Each Epoch and Specific Contribution of Top Domains

Fig re 1 also shows the extent amongst first-party servers for the top root domains in the Sep'08 epoch of our study. These domains were generally at or near the top of all epochs. Apart from google-analytics.com and quantcast.com, which were initially not present in data from early epochs, the other domains in Figure 1 were at or near the top in all epochs. These results show that beyond a general increase in the footprint of all domains, the footprint of some domain has grown significantly. The domain do bleclick.net had the largest penetration in the first epoch and has more than doubled its penetration since. The q antserve.com domain is only present in the latter two epochs, but is now one of the top few domains. The google-analytics.com domain was not present in our first epoch yet has leapfrogged to near the top over the co rse of our study.

4.2 What Are These Top Third-Party Domains Doing?

Given the spread of these third-party domains amongst first-party servers, it is important to nderstand what these third-party domains are doing. Originally, third-party cookies were sed to track sers, but techniques employing combinations of first-party cookies and JavaScript execution have also become common.

Rather than study all third-party domains, we focused on those with at least a one-percent penetration in a meas rement epoch. Using this list as a starting point, we studied traces of requested objects, cons lted the browser cookie database, and examined downloaded third-party JavaScript to better nderstand the nature of content served by servers in these domains. We primarily focused on the use of cookies by these third-party domains for tracking and whether these domains were using JavaScript to track sers in conj nction with se of first- or third-party cookies. We found fo r types of third-party domains that track sers amongst the set we examined.

1. Third-party domains that only set third-party cookies to track sers and do not make se of JavaScript for additional tracking. From Figure 1 these incl de doubleclick.net, atdmt.com and 2o7.net.

2. Third-party domains that se JavaScript with state saved in first-party cookies. A prominent domain of this type is google-analytics.com, which uses a piece of JavaScript code to interrogate the first-party cookies of the site and then retrieves an object sing an identifying URL for sending information back to its third-party server.

3. Third-party domains that use both third-party cookies and JavaScript to set first-party cookies. The domain quantserve.com is an example of such a third-party domain that use JavaScript as well as both first-and third-party cookies to track ser actions.

4. Third-party domains that do not se JavaScript for setting first-party cookies nor se third-party cookies. However these domains are involved by serving ads URLs with tracking information, such as adbrite.com or adb reau.net. Others are owned and operated by a third-party domain that does tracking. For example, instances of 2mdn.net virtually always occ r in conjunction with doubleclick.net.

Another potential means for third parties to track users is "Flash cookies," which are Local Shared Objects (LSOs) maintained by the Adobe Flash Player [9]. These LSOs are stored on a ser's computer in a local repository maintained by the Adobe Flash player. We examined results for our test data set to see if the presence of s ch third-party Flash cookies in the form of local shared object files could be detected. In the data we did observe one s ch instance where the Flash script file quant.swf was served by the server flash.quantserve.com with subsequent URL retrievals back to this third-party server. This Flash script is working similarly to one in JavaScript, but instead of saving state using cookies, it is using one of these LSOs to save state at the browser. Unfortunately, these cookies are not controlled via standard privacy settings of browsers so a ser may not be aware they are even set.

4.3 Company Acquisitions

Apart from the growth of individ al domains, acquisitions in the industry over the course of our study have changed the landscape and created families of companies that have multiple perspectives of user viewing habits. Table 1 shows a list of third-party acquisitions by third-party parent domains with a presence in at least 1% of first-party servers. The list was compiled by the authors sing information gleaned from public announcements.

Table 1: Known Acquisitions of Third-Party Domains By Parent Companies

Parent	Acquired	Date
AOL	advertising.com	Jun'04
	tacoda.net	Jul'07
	adsonar.com	Dec'07
Doubleclick	falkag.net	Mar'06
Google	youtube.com	Oct'06
	doubleclick.net	Mar'07
	feedburner.com	Jun'07
Microsoft	aquantive.com (atdmt.com)	May'07
Omniture	offermatica.com	Sep'07
	hitbox.com	Oct'07
Val eclick	mediaplex.com	Oct'01
	fastclick.net	Sep'05
Yahoo	overture.com	Dec'03
	yieldmanager.com	Apr'07
	adrevolver.com	Oct'07

Using the data of Table 1 we can follow the growth both in terms of internal expansion and external acquisitions for prominent third-party companies. In the following, the families are presented in order of the res lting size measured in terms of penetration within our set of first-party servers.

Fig re 2 shows the growth of the Google family of domains over the co rse of our study. Within each epoch, two sets of bars are shown. The right-most bar contains constituent members of Google at each epoch. Thus in Oct'05, the primary extent of Google was due to googlesyndication.com, although moving to Oct'06 google-analytics.com was niquely sed on more first-party servers with some sites having an overlap of more than one Google domain. Moving forward in time, the Google domains googleadservices.com, google.com and googleapis.com serve some third-party content.

The left-most bars in each domain show the extent of non-Google domains that are eventually acquired by Google. The most prominent is doubleclick.net, which includes 2mdn.net and the acquisition of falkag.net after Mar'06. After the acquisition of Doubleclick by Google in Mar'07 the extent of the Google family shows a sharp increase in our Feb'08 epoch. After the acquisition, doubleclick.net and google-analytics.com each contrib te significantly to reach of this family of domains with the large overlap primarily due to first-party servers employing both of these domains. The end result is that in the Sep'08 epoch, the Google family has a reach of nearly 60% amongst the set of domains in o r core test data set—the highest among all third parties by far.

Fig re 3 shows the growth of the Omniture family of domains. This family has grown thro gh the increase se

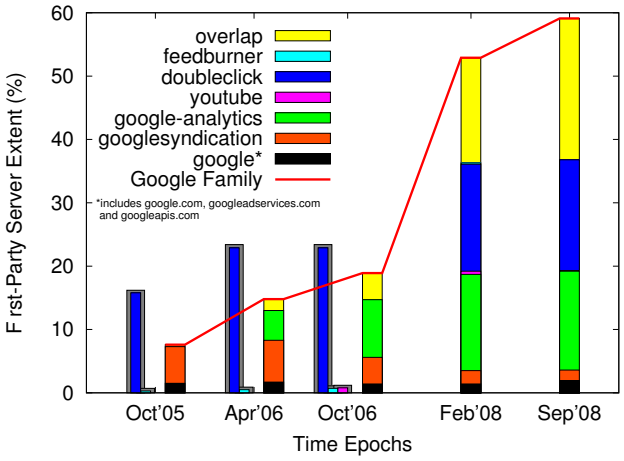

Figure 2: Growth of the Google Family

of Omniture third-party servers, primarily the 2o7.net domain, as well as the acquisition of the offermatica.com and hitbox.com domains. In Sep'08 the family has a reach of 28% with most of it due to the original omniture.com domain.

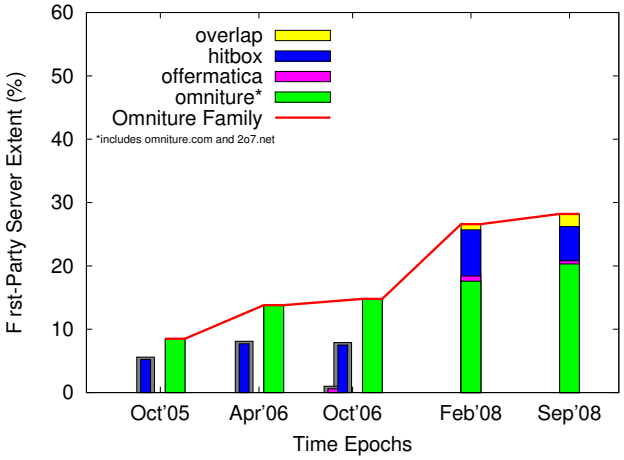

Figure 3: Growth of the Omniture Family

Figure 4 shows the growth of the Microsoft family over the course of o r study. This family of domains has a reach of 22% in Sep'08 with its growth due almost entirely to the acquisition of Aquantive and its atdmt.com domain.

Figures 5 and 6 track the final two significant families, Yahoo and AOL, over the course of our study. Yahoo has a reach of 15% in Sep'08 with much of its growth due to the acquisition of the yieldmanager.com domain. AOL has a reach of over 14% in Sep'08 due to two acquisitions in 2007 and its acquisition of advertising.com prior to the beginning of our study. Valueclick, the last family listed in Table 1, has a much smaller extent of 4% in Sep'08 and is not shown.

Once acquisitions are assigned to their respective parent company, Figure 7 takes a similar approach as Fig re 1 of first showing the extent of the top-10 family during each

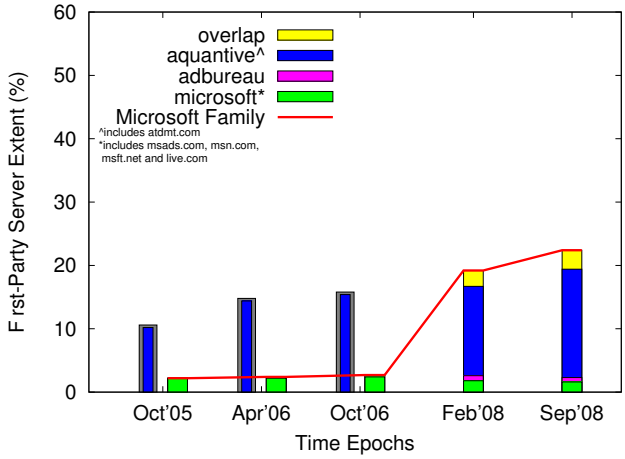

Figure 4: Growth of the Microsoft Family

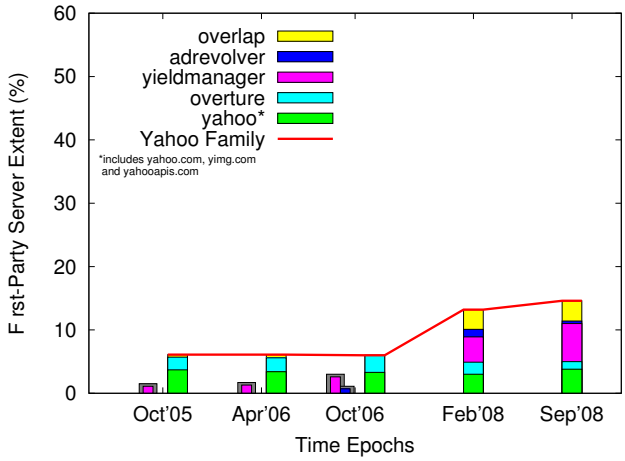

Fig re 5: Growth of the Yahoo Family

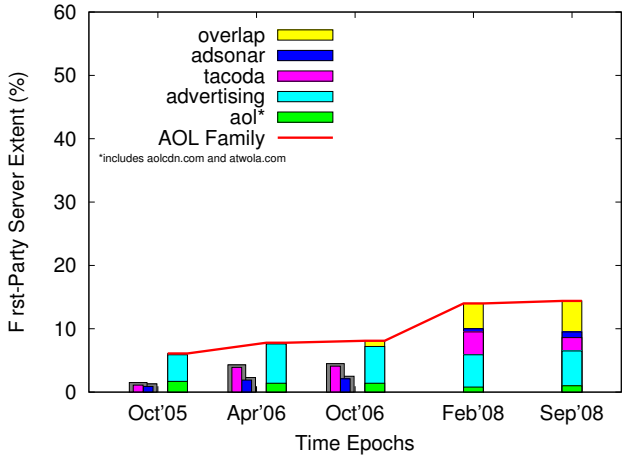

Fig re 6: Growth of the AOL Family

epoch and the top families for the Sep'08 epoch. Relative to Figure 1, 2mdn.net is merged into doubleclick.net, which is then merged into the Google family along with the domain google-analytics.com. Similarly, atdmt.com becomes part of the Microsoft family and 2o7.net part of the Omniture family. The res lts show that acquisitions have helped to create the five families of domains with highest penetration with quantserve.com and revsci.net being the two independent domains with the highest penetration.

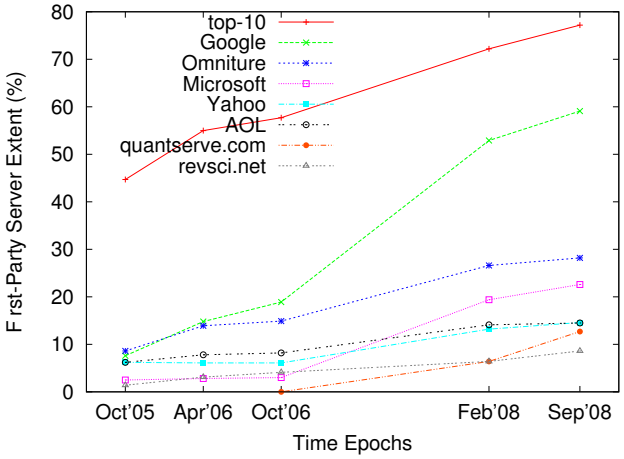

Figure 7: Extent of Top-10 Third-Party Families in Each Epoch and Specific Contrib tion of Top Families

4.4 Depth of Third-Party Penetration

Another way to understand the extent of third-party penetration is to examine the depth of these domains by determining how many independent families and domains are associated with each first-party server. For this analysis, we first assigned each root domain and then determined all families with at least a one-percent penetration for each epoch. We then analyzed the n mber of these top third-party families that are associated with each first-party server. Res lts of this analysis are shown in Table 2.

Table 2: Depth of Top Third-Party Penetration Amongst First-Party Servers (%)

Time	% 1st-Party Servers w/ No. Top 3rd-Party Domains				
Epoch	≥ 1	≥ 2	≥ 3	≥ 4	≥ 5
Oct'05	53	24	12	5	1
Apr'06	63	35	19	10	2
Oct'06	66	38	23	13	6
Feb'08	76	47	29	18	10
Sep'08	81	52	34	24	14

The res lts show that the percentage of first-party servers with m ltiple top third-party domains has risen from 24% in Oct'05 to 52% in Sep'08. This increase has occ rred despite the merger of previously independent domains thro gh acquisitions. This increase is significant because it shows that now for a majority of these first-party servers, sers are being tracked by two and more third-party entities.

4.5 Extent of Company Families in Consumer Sites

In addition to the broad set of popular sites we se in o r study, we also wanted to focus on cons mer sites which a large number of users are likely to visit in order to make p rchases rather than simply browse. These sites elicit more information about users who are less likely to be browsing anonymously as compared to, say, news Web sites. In order to make use of our longitudinal data we identified a subset of 127 test data set sites across the Alexa categories for examination in this portion of our study. Results for this s bset of consumer sites are shown in Figure 8.

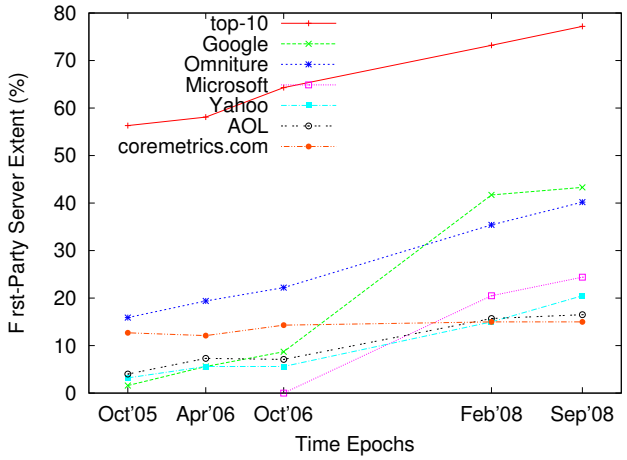

Fig re 8: Extent of Top-10 Third-Party Families in Each Epoch and Specific Contribution of Top Families for Cons mer Sites

The extent of top-10 third-party domains is comparable to Fig re 7, although there is variation in the extent of specific domains. The Google family is still the largest in Sep'08, but smaller than across all first-party servers while the Omniture family is larger for cons mer sites.

Also interesting is two third-party domains that are in the Sep'08 top-10 for cons mer sites. These sites were not shown in Figure 8 to reduce the vis al complexity of the graph. The domain abmr.net has a 6% extent in Sep'08. It is significant beca se it is owned by Akamai and tracks sers via third-party cookies. Given that in Sep'08 66% of first-party servers were sing Akamai's CDN service to directly serve first-party or indirectly serve third-party content, the introduction of a CDN-based tracking service has potential privacy impact. The presence of this domain, which was first observed in the Feb'08 epoch, coincides with a patent application from Akamai on data collection in a CDN [15].

Another domain with a 6% extent in Sep'08 is specificclick.net, domain for Specific Media. It was recently reported that Specific Media has created profiles on more than 175 million individual users [21]. Its higher presence in cons mer sites compared to the larger set of sites indicates that cons mer sites tend to be more val able for this type of profile tracking.

4.6 Extent of Company Families in Fiduciary Sites

We also examined another set of sets, originally used in [13]—Web sites involving the managing of personal fiduciary information. Users provide private information such as credit cards and bank account numbers to such sites. We used the 81 sites from [13] across nine categories: credit, financial, insurance, medical, mortgage, shopping, subscription, travel, and utility. Longitudinal results for these sites over three epochs are shown in Figure 9.

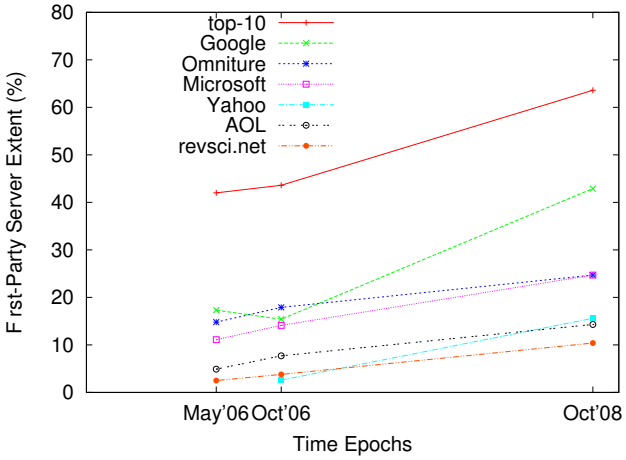

Figure 9: Extent of Top-10 Third-Party Families in Each Epoch and Specific Contribution of Top Families for Fiduciary Sites

The tone of these results is similar to what we found for the consumer sites although the extent of the top-10 third-party families in each epoch is a bit less. Given the increased privacy concerns that users have with sites such as those involving medical and financial concerns, the extents are still large.

4.7 Impact of Currency of Category Membership

Finally, we investigated the impact of changing membership in the Alexa categories used as the basis for our study. The membership of these categories was originally obtained in 2005 so an obvious question is whether the results change if we use current membership for the categories.

For this work we retrieved the membership of 15 Alexa categories [3] of popular sites in 2008. Twelve of these categories were in common with those we retrieved in 2005: arts, business, computers, games, health, home, news, recreation, reference, regional, science, and shopping. The 2005 membership of these twelve categories represented 1068 unique URLs while the 2008 membership represented 1111 unique URLs. Of these counts, there was an overlap of 625 URLs, thus nearly 60% of the URLs in 2005 were still popular in 2008.

The URLs for these twelve categories using the 2005 and 2008 memberships were each retrieved in Sep'08 and analyzed. The top-10 extent and the top families in Sep'08 are shown in Table 3 for the two membership periods.

The results show that despite the membership changes between the two time periods, the new membership results are consistent with the old with similar ordering and magnitude of the extent of the top-10 third-party domains. The extent of the third-party domains for the 2008 membership is consistently greater for the top third-party domains.

Table 3: Top Third-Party Family Extent Among First-Party Servers for 2008 and 2005 Period Memberships(%)

Third-Party Domain	Membership 2008	2005
top-10	79.4	78.7
Google	60.9	57.7
Omniture	33.8	30.0
Microsoft	24.4	22.7
Yahoo	15.8	14.9
AOL	15.6	14.8
quantserve.com	12.4	11.3
revsci.net	10.8	9.2

5. LIMITATIONS OF PROTECTION TECHNIQUES

Given the increasing penetration of third-party domains on popular Web sites, an obvious question is the effectiveness of potential actions that a user can take to protect against privacy diffusion. Prior work in [11] implemented and examined tradeoffs between effectiveness and page quality for a range of approaches with the best general approaches limiting the download of third-party content such as cookies, JavaScript and identifying URLs. The work found that restricting first-party content, cookies or JavaScript led to errors or sharper reductions in visual quality when downloading a page.

As a result, the obvious approach for a user interested in protecting their privacy is to not allow third-party cookies, which is a privacy option available in browsers; disallow third-party JavaScript execution through tools such as Firefox's NoScript extension [17]; and block known third-party identifying URL content using a tool such as Adblock Plus [1].

While each of these techniques does work, a careful analysis of how third-party aggregator sites are tracking users shows that all of these techniques are limited in their effectiveness for protecting users. Results of this analysis across the five time epochs of our study are shown in Figure 10 where third-party domain servers are increasingly "hiding" their content in first-party domain servers.

The first result is that third-party aggregators are not only using third-party cookies to track users as discussed in Section 4.2, but these aggregators are using first-party cookies to store information about a user's accesses to the first-party site. These first-party cookies are actually set by third-party JavaScript code such as urchin.js of google-analytics.com or Omniture's s_code.js. As shown in the FirstPartyCookies result of Figure 10 the percentage of first-party servers that have first-party cookies set and used by third-party JavaScript has grown to nearly 60% of all first-party servers over the course of our study. These first-party cookies are much harder for a user to not accept because doing so for all first-party cookies causes some first-party site access to break.

A related issue is the source of the JavaScript code that is used for tracking. One source is a third-party server, such as

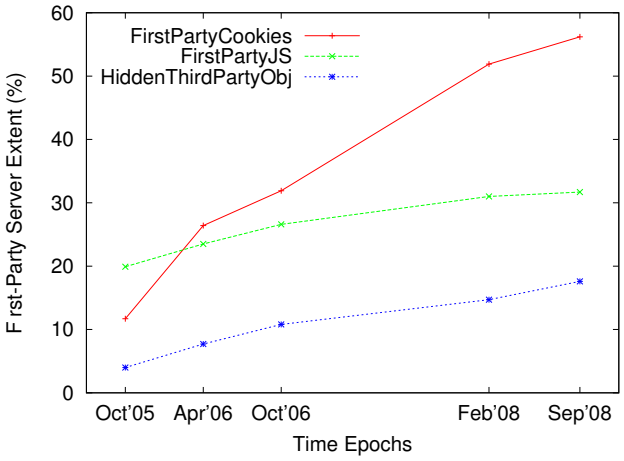

Figure 10: Growth of Hidden Third-Party Content

the case where urchin.js is typically served by the third-party server www.google-analytics.com. In such cases it is possible to use a URL blocker or NoScript extension to prevent download/exec tion of the code. Alternately, other third-party JavaScript code is actually served by a first-party server. For example on the first-party site abc.go.com, Omniture's JavaScript code is served by the server a.abc.com, which is a first-party server as confirmed by its ADNS. These cases are much harder to automatically block and as shown in the FirstPartyJS res lts in Figure 10 now occ r for over 30% of first-party servers. This figure is conservative as it is based on the extent of well-known names for tracking JavaScript code that we could identify in our data.

The outcome for exec ting one of these tracking JavaScript codes is the generation of an identifying URL that is "requested" from a third-party server in order to pass information back to the third-party domain. For example, urchin.js causes a 503-byte identifying URL to be sent to www.google-analytics.com in order to retrieve a 35-byte image. Again blocking s ch identifying URLs is possible when the URL is sent to a well-known third-party server, but increasingly this request is being sent to an apparent first-party server. For example, the Omniture JavaScript code on abc.go.com generates an identifying URL for the server w88.go.com, which is in the same domain as the first-party server, b t based on its ADNS is actually part of the Omniture network. Fig- re 10 shows that now close to 20% of first-party servers in our data set contain s ch third-party objects that are "hidden" in what look to be first-party servers.

The bottom line is that identifying and blocking third-party content sed for tracking is increasingly difficult as these third-parties work with first-party sites to place such content in servers that are or appear to be part of the first-party site. However these "first-party" servers are simply a DNS alias for what is actually a third-party server. This approach makes for limitations of c rrent tools that protect based on URL or server name to accurately identify what content to block. This is a similar "cat and mo se" game as we discussed in previous work on ads [12].

This game is also not limited to third-party sites doing analytics. Third-party sites doing behavioral tracking could deploy their content on what appear to be first-party domain servers and make se of first-party cookies to track sers across first-party sites witho t any apparent use of third-party content or cookies. While we saw little evidence of this approach in o r data, we would expect approaches like this to be sed if enough sers stop allowing third-party cookies.

Additional privacy protection tools are being made available in browsers. Microsoft has announced its InPrivate mode for IE8 [2] and Google has a similar "incognito" mode in its new Chrome browser [8]—this was originally available on Macintoshes. In each case, when a user invokes these modes then the browser does not save the ser's browsing history, cookies and other data. These tools are directed at "over-the-sho lder privacy" from others with access to the computer rather than protecting privacy from third parties as the capability to block cookies is already available. InPrivate does have capabilities to establish a favorites list for preservation of cookies, b t this feature requires active management of sites to add or the need to switch in and out of the InPrivate mode. The mode also automatically detects when a user has been "seen" by more than ten third-party sites, b t as o r results show detection of a third-party cannot always be done by string matching alone so the val e of this feature is not clear.

6. DISCUSSION

So far in this paper, we have examined well-known techniques for gathering privacy related information about sers and the degree of penetration in popular Web sites. We have also examined the role of cookies and JavaScript as well as the potential for blocking diff sion of private information. The examination of acquisitions of companies points to the potential of significant growth in aggregate data. In partic lar, the acquiring company has older data that they could not have otherwise obtained. By p rchasing behavioral data from the past, the acquiring company is able to get a broader idea about the behavior of sers over time which can be helpful to predict f ture trends. The ability to link (or "fuse") s ch data with other information heightens the risk of converting user-ne tral data into personally identifiable information. Our work has examined diff sion of private information at the level of Web site access. Most of this information happens relatively transparently altho gh sers may be aware of presence of cookies. The use of cookies (especially third-party cookies) and extraction of information via JavaScript is generally opaque to users.

Beyond the sites they visit, there is a great deal of private information that sers s pply to many Web sites. We examine a broad subset of these with a view of how data fusion could occur between the private information collection that we have examined th s far.

Search engines typically record the search strings entered by users and some search sites even make the history of past searches available to the user. Ask.com has a feature to erase the past searches. Rare exceptions like the new cuil.com search site explicitly indicate that no information about users is gathered or maintained [4]. However, most search sites can and do record information s pplied by users.

The problem gets a bit more complex when we examine the popular free Web email services. These services require sers to acknowledge that they accept a Terms of Services agreement, which spells o t how a ser's private information

will be treated. The social graph of a user can be constructed simply by mining the set of their comm nicants.

Toolbars are another potential so rce of privacy leakage. For example, MSN and Yahoo have toolbars available for Internet Explorer with optional features to help these companies to provide better service by sending information abo t visited URLs to these sites. The Google toolbar (which comes pre-installed on any Dell PC [18]) has a feature showing the page rank of each page visited by a ser. This rank is determined via a request, with attached cookie, to Google for each URL visited by a user.

As previously discussed, Google's new browser Chrome has an Incognito privacy feature, but has other features that raise privacy concerns [5]. All partial URLs or queries typed into Chrome are sent (by default) to Google and completion suggestions are generated. Thus, Google can record the list of URLs users attempt to visit even if there is no link between these Web sites and Google. The retention policy for these data is not specified in the browser's privacy policy.

Another potential source to gather information is online social networks (OSNs). One of these, (orkut.com), is part of the Google family of domains. In addition, we fo nd in [14] that the third-party domains found in popular Web sites are also prominent in the popular OSNs that we studied.

The top few family of domains that we disc ssed in Section 4.3, also operate search and free email services (AOL, Google, Microsoft and Yahoo) and deliver cookies as part of these services. Th s the potential for combining information available to them from registered sers clearly exists—for example linking the information available from any of these services with data aggregated from Web traversals. At the minimum, behavioral marketing introduces what has been termed the "creepiness factor" [20] where sers see ads targeted not j st on books that are bought, but on medical conditions that are looked p.

7. SECONDARY PRIVACY DAMAGE

One of the new iss es we are concerned about that does not appear to have been raised in the privacy literature is that of secondary privacy diff sion. In all the diffusion we have discussed thus far, the affected person is the one browsing the Web. The notion of secondary leakages arises when privacy related to other users are either deliberately or inadvertently leaked. Even if the original ser is libertarian and does not mind their private information leaking, they should not be contrib ting to diff sion of other people's privacy. We give examples of this phenomenon here.

Earlier in Section 6, we referred to the potential construction of social graph by Web-based email services. Witho t the recipient's knowledge or consent, the communication between the first ser (someone who has acceded to the Terms of Service) is available to the email service. If the recipient replies to the email then the contents of the response are also available without the second user ever being aware of the privacy policy of the email service.

Some Internet services allow c stomers to provide email addresses of other Internet users so that these other users can be invited to an event or to send copies of restricted online articles to non-s bscribers. Event organizing sites host content of interest to the event which can be pdated by the invited parties. However, the s pplied addresses become known to the service without any prior approval necessarily obtained from these other Internet users res lting in secondary leakage. The relationship between the s pplier of the email address and non-subscribers can be stored by the article site. For example, the forwarding of a news article of restricted sites to someone else may give an indication of the recipient's interest or political leanings.

Sites that allow tagging of pictures may store information about named users. The user-generated tags create linkages aro nd the content of the picture or may provide other relationship information (e.g. parent, sibling, etc) between sers.

Currently, there is no way to prevent secondary leakage before it occurs. However, if there is information about sers who were unaware of their information leaked by others without their knowledge or consent, then monitoring so rces of public information (public Web sites, social network pages, blogs) can help identify s ch leakage post facto. Such detection can lead to the user being notified and the ser can decide if s ch privacy leakage is acceptable. If not, the party that is the source of s ch p blic information can be notified to prevent f ture leakage.

It should be noted that the same aggregators who track the movement of sers across the Web can also gather available information abo t other users.

8. CONCLUSIONS

In this work we used o r long-term data to present a longitudinal analysis of privacy diff sion on the Web. This is the first study to measure this diffusion over an extended period of time. The res lts from the study show that penetration of the top-10 third-party servers tracking user viewing habits across a large set of popular Web sites has grown from 40% in Oct'05 to 70% in Sep'08.

During the same time period of this increased privacy diffusion, we observe a number of family of domains that have been created through acquisitions of one company by another. These acquisitions have decreased the n mber of popular independent third-party domains. The overall share of the top-five families: Google, Omniture, Microsoft, Yahoo and AOL extends to more than 75% of o r core test set with Google alone having a penetration of nearly 60%.

Not only are these families and other third-party domains represented broadly across our set of first-party sites, but the depth of this representation has increased to the point that in Sep'08 a majority of o r first-party sites made se of two or more third-parties. This result is significant because it shows users are being tracked by multiple entities when accessing a first-party site.

Finally we found that existing privacy protection techniques have limitations in preventing privacy diffusion. These techniques work by restricting the download of third-party content in the form of cookies, JavaScript and identifying URLs, b t o r res lts show that increasingly third-party aggregators are working to hide their presence in a first-party site by serving content from what are or appear to be first-party servers. This approach makes it diffic lt for tools that protect based on URL or server name and will likely increase in se as more users deploy privacy protection techniques.

The aggregation of tracking data, particularly by the families we identify, is of concern because of the other so rces of ser data that these families have available to them. Search terms, email services and toolbars are only some of the additional sources of information about sers available to families s ch as AOL, Google, Microsoft and Yahoo that be

linked with tracking data. Services such as email and social networking sites are also opportunities for secondary privacy leakage where private information about a user is made available to the service or public without the consent of the user.

Future work includes continuing to monitor the presence and activities of third-party aggregators. We have seen approaches evolve and expect that they will continue to evolve as there is a cat and mouse game between users interested in privacy protection and companies interested in gathering data. We plan to continue examining the relationship between tracking data and whether it can be fused with PII. Finally we plan to further examine the extent of secondary privacy leakage along with measures to limit its impact.

9. ACKNOWLEDGMENTS

We thank Trevor Jim and other anonymous reviewers for their comments on earlier versions of the paper.

10. REFERENCES

[1] Adblock plus: Save your time and traffic. http://adblockplus.org/.

[2] C. Albanesius. Microsoft tips IE8 privacy features, August 26 2008. http://www.pcmag.com/article2/0,2817,2328900,00.asp.

[3] Alexa: Most popular web sites. http://www.alexa.com/.

[4] Cuil - your privacy. http://www.cuil.com/privacy/.

[5] W. Davis. Polish on google's new chrome tarnished by privacy questions, September 2, 2008. http://www.mediapost.com/publications/index.cfm?fa=Articles.showArticle&art_aid=89743.

[6] S. DeDeo. Pagestats, May 2006. http://www.cs.wpi.edu/~cew/pagestats/.

[7] European union directive 95/46/EC. http://www.cdt.org/privacy/eudirective/EU_Directive.html, Nov. 1995.

[8] A. Greenberg. Going 'incongnito' can you really web browse on the down low?, September 5, 2008. http://www.newsweek.com/id/157293?tid=relatedcl.

[9] I'm A Super.Com. Flash cookies: The silent privacy killer, October 9 2008. http://www.imasuper.com/66/technology/flash-cookies-the-silent-privacy-killer/.

[10] C. Jackson, A. Bortz, D. Boneh, and J. C. Mitchell. Protecting browser state from web privacy attacks. In Proceedings of the International World Wide Web Conference, Edinburgh, Scotland, May 2006.

[11] B. Krishnamurthy, D. Malandrino, and C. E. Wills. Measuring privacy loss and the impact of privacy protection in web browsing. In Proceedings of the Symposium on Usable Privacy and Security, pages 52–63, Pittsburgh, PA USA, July 2007.

[12] B. Krishnamurthy and C. Wills. Cat and mouse: Content delivery tradeoffs in web access. In Proceedings of the International World Wide Web Conference, Edinburgh, Scotland, May 2006.

[13] B. Krishnamurthy and C. E. Wills. Generating a privacy footprint on the Internet. In Proceedings of IMC, October 2006.

[14] B. Krishnamurthy and C. 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.

[15] Michael Afergan and Thomson Leighton and Timothy Johnson and Brian Mancuso and Ken Iwamoto. Method of data collection among participating content providers in a distributed network, 2008. United States Patent Application 20080092058. http://www.freepatentsonline.com/y2008/0092058.html.

[16] Mike On Ads. How do behavioral networks work?, February 28 2007. http://www.mikeonads.com/2007/02/28/how-do-behavioral-networks-work/.

[17] Noscript. https://addons.mozilla.org/firefox/722/.

[18] S. Olsen and T. Krazit. Dell embraces google, May 25, 2006. http://news.cnet.com/Dell-embraces-Google/2100-1032_3-6077051.html.

[19] Pests clasified in the category tracking cookie. http://www.pestpatrol.com/zks/pestinfo/tracking_cookie.asp.

[20] Privacy on the web: Is it a losing battle?, June 25, 2008. Published in Knowledge@Wharton. http://knowledge.wharton.upenn.edu/article.cfm?articleid=1999.

[21] P. Whoriskey. Candidates' web sites get to know the voters presidential campaigns tailor, target ads based on visitors' online habits, August 30 2008. http://www.washingtonpost.com/wp-dyn/content/article/2008/08/29/AR2008082903178_pf.html.

[22] Consumer tips: How to opt-out of cookies that track you. http://www.worldprivacyforum.org/cookieoptout.html.