Privacy leakage vs. Protection measures: the growing disconnect

Balachander Krishnamurthy
AT&T Labs–Research
bala@research.att.com

Konstantin Naryshkin
Worcester Polytechnic Institute
konary@wpi.edu

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

ABSTRACT

Numerous research papers have listed different vectors of personally identifiable information leaking via traditional and mobile Online Social Networks (OSNs) and highlighted the ongoing aggregation of data about users visiting popular Web sites. We argue that the landscape is worsening and existing proposals (including the recent U.S. Federal Trade Commission's report) do not address several key issues. We examined over 100 popular *non-OSN* Web sites across a number of categories where tens of millions of users representing diverse demographics have accounts, to see if these sites leak private information to prominent aggregators. Our results raise considerable concerns: we see leakage in sites for every category we examined; fully 56% of the sites directly leak pieces of private information with this result growing to 75% if we also include leakage of a site userid. Sensitive search strings sent to healthcare Web sites and travel itineraries on flight reservation sites are leaked in 9 of the top 10 sites studied for each category. The community needs a clear understanding of the shortcomings of existing privacy protection measures and the new proposals. The growing disconnect between the protection measures and increasing leakage and linkage suggests that we need to move beyond the losing battle with aggregators and examine what roles first-party sites can play in protecting privacy of their users.

1. INTRODUCTION

Recently, multiple vectors of private information leakage via Online Social Networks (OSN) and the two-decade long aggregation of data about users visiting popular Web sites have been reported. The problem of privacy has worsened significantly in spite of the various proposals and reports by researchers, government agencies, and privacy advocates. The ability of advertisers and third-party aggregators to collect a vast amount of increasingly personal information about users who visit various Web sites has been steadily growing. Numerous stories have expressed alarm about the situation with legislatures and privacy commissioners in different countries paying closer attention to the problem [14]. The awareness about the steady erosion of privacy on the part of users is growing slowly. The potential economic impact as a result of loss of brand value has forced some companies to start paying closer attention to complaints from users and privacy advocates.

In this paper we argue that the privacy landscape is worsening as there is a *growing* disconnect between steadily increasing leakage to and linkage by aggregators with existing and proposed protection measures. We show that beyond the egregious leakage of private information via OSNs and their more recent mobile counterparts, a key part of the Internet with tens of millions of users representing diverse demographics with accounts on popular *non-OSN* Web sites also suffer from private information leakage to prominent aggregators. Additionally, less well-understood notions of *linkage* are typically not addressed by most of the proposed privacy solutions. One such privacy issue arises from the existence of globally unique ids such as an OSN id or reused email addresses that could be used to link together pieces of seemingly distinct information. Beyond the intrinsic identifying nature of these ids, they aid in linking together other information, such as cookies from a home and work computer. New proposals, such as the recent United States Federal Trade Commission's December 2010 report [10], fail to address several key issues.

Our earlier work focused on longitudinal data gathering by aggregators on the Web [15], leakage of personal information via popular OSNs [13] and the more recently mobile OSNs [16]. However, there has been no attention paid thus far to another segment of the Internet where sites encourage and allow users to create accounts so that they could have a richer interaction experience. Many popular Web sites allowed users to establish profiles long even before the advent of OSNs. There are significant demographics that are present in non-OSN Web sites that may not be on OSN sites and their private information is also of interest to aggregators. On many of these sites, users create profiles with varying amounts of personal information, but typically less than what they supply on OSN sites. Unlike OSNs, these Web sites already have content and do not depend on users to create content; users could however add comments or tags. Surprisingly, there is considerable overlap in the nature of personal information that users provide across these sites. We should also note that the degree of sensitivity to different aspects of their personal information varies across users as is the potential for identifiability (ability to link a unit of personal information with a specific user).

We look at a broad array of sites in various categories

where users establish identities and provide personal information. We examine the extent of *direct* leakage of private information as a result of typical user actions on these sites and present a view of exactly what subset of private information that third-party aggregators receive from these Web sites. Finally, we explore the potential for aggregators to link various pieces of information they receive via globally unique identifiers, such as userids from these sites, or via browser fingerprinting.

Note, we take a user's perspective in labeling any private information transmitted from a first-party site to a third-party server as "leakage". In some cases, a first-party site may knowingly transmit such information and have contractual agreements with third parties preventing potential linkage of information obtained from different sites. Many first-party sites outsource analytics work to third parties and share user information with them to obtain tailored analytics. The private contract between a first-party site and such third parties provides a degree of safety that the third parties will treat data received through such transfers as belonging to that of the first party and not use it for other purposes, such as linking with data received from other sources. Disclosure of such contracts would lessen some of the concerns that we raise in this paper.

We also show how a new consumer privacy protection proposal by the U.S. Federal Trade Commission falls short in dealing with several key privacy related issues. The privacy community needs a clear understanding of the shortcomings of privacy protection proposals and the need to overcome the structural difficulties. In the ongoing cat and mouse game between users and aggregators, the economic advantage is on the side of aggregators. Rather than fight a losing battle with aggregators, we need to examine what roles *first-party* sites can play in protecting privacy of their users: a hitherto unexplored avenue.

The rest of this paper is organized as follows: Section 2 describes our study of leakage of sensitive private information from popular non-OSN Web sites to and possible linkages by third-party aggregators. Section 3 discusses the results of our study. Using results from this and earlier work, Section 4 examines the shortcomings of existing privacy protection measures and new proposals in combating leakage and linkage of private information. We conclude with a summary of our findings and directions for future work.

2. STUDY

Our study focuses on the leakage and potential linkage of *private information* enabled by first-party sites to third parties. We examine a broad range of Web sites where significant numbers of users register and supply personal information while setting up an account. We initially look at the degree of leakage of private information via different sites focusing on *direct* leakage of "bits" of private information (e.g., name, email address, and gender) to third-party aggregators. We begin by defining a methodology for identifying categories of sites and determining a specific set of sites to study within each category. We describe how we collected data for each site and how the data were analyzed both for leakage of private information and for potential linkage of this information by third parties.

2.1 Categories and Sites for Study

Users are encouraged to create accounts for many categories of sites and it is often a pre-requisite for users in job-related or dating sites. Other categories allow registered users to upload content while restricting unregistered users to browse content. Registered users can post reviews and comments, personalize the site, participate in contests, save purchase information, receive electronic newsletters, and gain access to restricted site content. Although only a minority of users may value these functions to register, for high-volume sites even a small fraction represents a large number of users.

We used Alexa (www.alexa.com) categories and subcategories including sites with a significant number of registered users that allowed for registration without any need for credit card information. Using self-reported numbers on the site or in published reports about the site we set a threshold of a minimum of 100,000 registered users (on most sites this number was in the millions). For sites that we were unable to find registration numbers, we included them if it was a popular site in a category where other sites had evidence of significant registration numbers. We also required that sites in a category be consistent in terms of why users register and what features registration provides.

To ensure we had the most popular sites, we began with the top-ranked site in an Alexa category (or subcategory) and worked down the list checking for the above criteria until we reached a target of ten sites within the category. Categories where we were unable to obtain ten sites were dropped. We identified ten categories and sub-categories for study (out of the 17 Alexa categories): Arts, Employment, Video Game News, Photo Sharing, News, Travel, Shopping, Relationships, Generations and Age Groups, and Sports.

Additionally, we examined two other Alexa categories. The first is Online Social Networking (OSN)—studied earlier in [13]—but included because of the huge numbers of registered users for sites in the category and to provide a way to contrast OSN and non-OSN Web sites. The second category is Health since users often supply potentially sensitive information to such sites. The search terms used or pages viewed could indicate interest in a sensitive medical condition and availability of such information to third parties could result in link-

ing it with other private information about a user. We used a similar methodology for determining ten sites in this category, although we relaxed the requirement of needing user registration as private information could be leaked from these sites even without explicit user identification. We established an account if registration was possible for these sites.

2.2 Data gathering methodology

We captured HTTP requests and responses from our Web browser using the Fiddler [9] Web proxy and examined it for visibly transmitted private information. Encrypted information or information transmitted over SSL could not be observed albeit for a tiny fraction.

The initial steps for testing each site consisted of creating an account, confirming a verification email message if needed, and viewing/editing the user profile on the site. A number of sites in our study allow a user to create an account and sign in via an existing third-party account, such as Facebook, Google, or Twitter. In cases where we could establish an account directly with the first-party site, we always chose that option. Some sites provide an opportunity to "remember me" on login, which was selected when available as some sites may then store private information (for example, in cookies) and subsequently leak it to a third party.

The set of actions tested for each site varied with its category and were tailored to the site's available features. Where feasible, we included actions that exercised features *only* available to registered users. In some categories (e.g., Arts and News) we included actions available to all users. For many of the sites there were a set of common actions: browsing, appropriate searches, and posting comments or reviews on the site's content. Beyond category-specific actions, many sites also provide opportunities to share content with friends via email and connections with social networking sites. When available, we shared content with sites in our OSN category and emailed articles to "friends".

We searched the gathered HTTP request/response (and POST) data for each site for leakage of userids, usernames, and pieces of private information to third-party servers. We generated a set of strings extracted from a user's profile that might be leaked to a third party. The set included all strings that users are mandated to enter into their profile at account creation time, such as email addresses, names, and zip codes. We also included other strings that users typically add to their profile that might be sensitive. Beyond profile data, search queries sent to Health and Travel (in the form of travel dates and cities) Web sites were added to the collection of search strings.

We processed the resulting data by systematically looking for leakage in the HTTP headers and eliminated false positive matches by hand (e.g., zip code being string present as part of a longer string). When leakage was seen, we recorded the leaked information, manner of leakage, and the third-party recipient(s). It is important to note that we only report *observed* leakage; i.e., our results are a *lower bound* on the extent of leakage. We may not observe leakage to a third-party for a number of reasons: we did not test an action available on a site that leads to leakage, the information is sent in a form we do not detect or is encrypted by a script before being sent to the third-party, or the transmission is encrypted (sent over SSL).

3. RESULTS

3.1 Leakage Results

We show examples of leakage of information to third parties after enumerating common actions for registered users of sites. We present leakage across site categories and conclude with an examination of the sensitivity and identifiability of the bits leaked to third parties.

3.1.1 Interaction with Sites

We enumerate a series of actions that users perform on sites including mandatory (such as creating and logging in to an account) and popular actions (such as editing profiles, searching etc.). In some cases, an interaction might be a sequence of actions. We present actual examples highlighting when private information is leaked to an aggregator[1] with the total numbers of sites leaking information shown in the following section. All data were originally gathered in September/October 2010 with all examples shown re-confirmed in April 2011. The confirmation found all examples of leakage continue with a few changes in the third-party recipient of the leaked information.

1. Account Creation/Confirmation: The first step is creating an account, which for some sites requires responding to an account confirmation email. We observed leakage of private information during this process in a handful of sites when the information is transmitted as part of the Request-URI of a HTTP GET request and then this Request-URI is contained in the `Referer` header for subsequent requests of embedded objects from third parties. Figure 1 shows an example of such leakage where a user's email address (private information is shown in **bold font** in examples) is leaked via a Sports category website as part of a HTTP request to a doubleclick.net server.

2. Account Login and Navigation: Upon login, some sites store private information about the user, such as name or email address, in site-specific, first-party cookies. Leakage of this private information occurs when

[1]We have masked the names of the first-party sites, but have notified them so that they can make the necessary changes.

GET http://ad.doubleclick.net/adj/...
Referer: http://submit.SPORTS.com/...?email=**jdoe@email.com**
Cookie: id=35c192bcfe0000b1...

Figure 1: Email Leakage in Account Confirmation

these sites also employ what we refer to as *hidden third-party* servers where a given server appears to belong to a first-party domain, but actually belongs to a third-party [15]. An example of this type of leakage is illustrated in Figure 2 where email, full name and zip code are passed to a URL belonging to a Age Groups category site because the cookies containing these values are associated with the Age Groups category site's domain and the browser interprets this server as being from the same domain. However examining the authoritative DNS server associated with this server shows that it actually belongs to the third-party domain 2o7.net (owned by Adobe), and this third-party is being leaked this information via this Age Groups category site.

GET http://metrics.AGEGROUPS.site/b/ss/..global/...
Referer: http://www.AGEGROUPS.site/
Cookie: ...e=**jdoe@email.com**&f=**John**&l=**Doe**&...&p=**12201**...

Figure 2: Email, Name and Zip Code Leakage Via First-Party Cookies to Hidden Third Party

We also observe leakage of information to a third-party server via the Request-URI when a user has logged into a site. The actions can be as benign as listening to a collection of songs from a music site or viewing a sequence of videos. Figure 3 shows an example of such leakage where a user's gender, zip code and music interests are leaked directly to doubleclick.net via an Arts category site, when playing songs.

GET http://ad.doubleclick.net/adj/...radio;ag=**30**;
gnd=**1**;zip=**12201**;artist=R53599;genre=**rock**;...
Referer: http://www.ARTS.com/...
Cookie: id=35c192bcfe0000b1...

Figure 3: Gender, Zip, and Interests Leakage in Navigation

3. Viewing/Editing User Profile: Once an account is created, a typical action for a newly registered user on a site is to view and edit the user's profile page. Some sites show information about a user in the title of this profile page, which is then leaked to third parties present on the page that are executing JavaScript code to collect information about the page. Figure 4 shows an example from the profile page of a user on an Arts category site, where the user's full name is leaked to a scorecardresearch.com server because JavaScript code executing in the browser obtains the title of the current page "ARTS - John Doe's profile" and passes it as an argument in the Request-URI. The user's Web site userid is *also* leaked via the `Referer` header.

GET http://beacon.scorecardresearch.com/...
c8=ARTS - **John Doe**'s profile...
Referer: http://www.ARTS.com/profile/public/|**123456789**...

Figure 4: Full Name Leakage Via Page Title

4. Inputting Content: Sites typically provide a means for a registered user to input content (e.g., for their profile). User's input is often sent to the server site by including it as parameters in a HTTP GET request to the server instead of using POST. However, if the page contains embedded third-party objects then the retrieval of these objects passes the user input to the third party in the `Referer` header. Figure 5 shows an example of such leakage, where age, zip code and gender information about a user of a Photo Sharing site, are leaked to a specificlick.net server.

GET http://afe.specificclick.net/?l=7654&sz=200x250...
Referer: http://a.PHOTOSHARING.com/hserver/
age=**30**/zip=**12201**/gender=**M**/...

Figure 5: Age, Zip and Gender Leakage Via Input

5. Searching for Sensitive Terms: Search terms are highly sensitive in some categories (e.g., Health) where users expect them to stay entirely within the site. Figure 6 shows an example where the search term "pancreatic cancer" is sent to a quantserve.com server via a site in the Health category.

GET http://pixel.quantserve.com/pixel;r=1423312787...
Referer: http://search.HEALTH.com/search.jsp?q=**pancreatic+cancer**

Figure 6: Sensitive Search Term Leakage Via Health Site

Figure 7 shows a similar example from a Travel category site, where a user's search term is a travel itinerary traveling from BOS (Boston) to MCO (Orlando) on specific travel dates. We see that this information has already been leaked to a doubleclick.net server and this server is now "daisy chaining" [11] this request (along with leakage of private information) to a revsci.net server.

3.1.2 Privacy Leakage Across Categories

Given the above actions and observed leakage, Table 1 shows the count of 10 first-party sites in each category directly leaking private information to at least one third-party for each action. Categories are listed in order of ones with the most number of sites leaking information to ones with the least number of sites with direct leakage. Travel and Health are at the top primarily because there is much leakage of search terms. The majority of OSN sites do leak information directly, but the Employment and Arts categories have at least as many sites exhibiting direct leakage. Fortunately, Age Groups category sites (mostly geared towards youngsters) exhibit the least amount of direct leakage.

GET http://pix04.revsci.net/...TRAVELSITE
Referer: http://fls.doubleclick.net/...u11=Economy/Coach;u3=**BOS**;
u4=**20110415**| **20110417**;u1=Flight;u2=**MCO**...

Figure 7: Itinerary Leakage Via Travel Site

Table 1: Leakage of Personal Information Via Web Sites Across Categories

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

The last row in Table 1 shows cumulative results for each column. These results show that 67 (56%) of the 120 sites across all 12 categories in our study directly leak private information to at least one third-party.

Counts for the various types show that leakage during account creation is infrequent while leakage of user information once a user logs in, such as shown in Figures 2 and 3, occurs for 27 sites across all categories. Leakage of sensitive search terms is prevalent in the Health and Travel categories. Although not a focus of this study and not reported in Table 1, we also observed that 58 (48%) sites leaked a userid for a site to a third-party as was noted in the example of Figure 4. Leakage of a site userid continues to be widespread in OSNs as previously reported in [13]. We look at how they could be used by aggregators to link information in Section 3.2.1. A total of 90 (75%) sites in our study directly leaked private information or a userid to a third-party.

3.1.3 *Sensitivity and Identifiability of Leaked Bits*

Table 1 shows that bits are being leaked by a number of sites. However these counts do not consider the *significance* of these leaked bits. We now examine privacy leakage along two axes: sensitivity and identifiability. Initially we assign a measure (high, medium or low) of sensitivity and identifiability to the privacy bits that users tend to disclose in various Web sites. Figure 8 shows our assignment of degrees of sensitivity and identifiability for various bits of personal information that we observe being leaked by our 120 first-party sites to a third-party. Other bits, such as cellphone number, credit card, SSN, DOB, mother's maidenname, IP address, photo and sexual orientation, are not directly leaked by any sites. An IP address is associated with each request, but is not directly a property of the request and we discuss this later in Section 3.2. A photo is commonly available in a user's public profile, but is not directly leaked to a third party. A user's cellphone number, DOB and orientation appear in a small number of public profiles, but are not directly leaked.

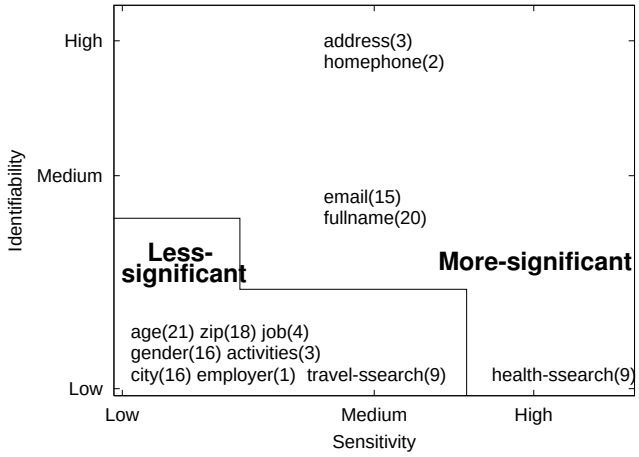

Figure 8: Sensitivity and Identifiability of Leaked Bits

The count next to the bits shown in Figure 8 are the number of sites (out of 120) directly leaking the information to at least one third-party aggregator. The age bit count also includes instances of "year of birth" leakage while "job" includes "occupation" and "career". There are two sensitive search terms in Figure 8. "Health-ssterm", such as a medical condition or a drug, are sent to Health sites. This information is leaked to at least one third party by *9* of the *10* sites in this category. "Travel-ssearch" is a travel search term representing travel cities and dates used in booking airline travel—also leaked by 9 of the 10 sites in the Travel category. These bits are low in identifiability, but higher on the sensitivity scale and are of concern if they can be linked to a user's identity. We suspect that this disclosure would come as new and unwelcome news to most users.

In terms of bits that can be used to identify users, Figure 8 shows that address, home phone, email address and full name are directly leaked by a number of sites. A user could be identified with any of these pieces of information[2]. The remaining bits in Figure 8 are low on the sensitivity and identifiability scales, but could be used to help identify a user if linked with other information. Information such as age, zip code, city and gender are made available to third parties via many sites.

3.2 Linkage Results

[2]http://www.time.com/time/business/article/0,8599,2058114,00.html

Beyond direct leakage of private information, we examined how seemingly disparate pieces of information can be linked together by aggregators. This linkage can be primarily done through unique identifiers attached to some of these records. Uniquely assigned cookies are traditionally used by third parties for such linkage, but as shown below, other identifiers and methods can be used even in the absence of cookies.

3.2.1 Linking Records Using Globally Unique Ids

Many third parties employ third-party cookies to link together records of information that they receive from a single browser. Now suppose a privacy-conscious user periodically removes the cookies stored in their browser or that a user employs separate home and work computers to access the same set of Web sites. In these cases, third parties will not receive the same cookie value for all accesses by the same user. However, if an aggregator is able to receive some type of globally unique identifier (GUID) for a user then the aggregator is in a position to link together the apparently disparate records associated with two separate third-party cookies.

When the userid (typically numeric when assigned by the site or alphanumeric when chosen by users) for a first-party site is combined with the first-party site from which it originates, it becomes the GUID for a *single* user. If the user consistently uses a first-party site that causes leakage of a unique id to a third-party then this unique id can be used to link together records associated with separate third-party cookie values.

In looking at the scope of this potential privacy issue, we reported in Section 3.1.2 that 48% of the sites studied leaked a userid for a site to a third-party. Beyond site userids, most sites require an email address for registration, which is another GUID (assuming a user uses the same email address across sites). When leaked, this address allows a third party to both personally identify the associated user and link together records containing it. Figure 8 shows that 15 (13%) sites leak an email address to a third party. Figure 9 shows an example of leakage of private information via a Employment category site, to a doubleclick.net server (daisy chained from a nexac.com server).

GET http://ad.doubleclick.net/activity;...
Referer: http://f.nexac.com/...http://www.EMPLOYMENT.com/...
na_fn=**John**&na_ln=**Doe**&na_zc=**12201**&
na_cy=**Albany**&na_st=NY&na_a1=**24 Main St.**&
na_em=**jdoe@email.com**...
Cookie: id=22a348d29e12001d...

Figure 9: Potential Linkage of Records Using Email Address

While not only leaking significant private information about the user to doubleclick.net, this leakage includes the user's email address. This is the same email address that doubleclick.net received in the example of Figure 1 except with a different cookie value thus allowing the aggregator to link these records. Now, if we look at the example of Figure 3, it has the same cookie value as the one in Figure 1 which was just linked above. Thus, the various bits of private information leaked in Figure 3 can also be merged by the aggregator. In addition, names and email addresses allow third-party linkage with public records of information [20].

We also saw a few cases where one site leaks the identifier of a user on a *different site*. For example, when a user on two different News sites shares a story with their Facebook account, that user's Facebook userid is stored in the respective first-party site's cookies and later leaked via these sites to hidden third parties. This leaked Facebook userid is a GUID and can be used to link together records received by the third parties.

3.2.2 Linking Records Without Cookies

Many, but not all, third-party aggregators employ third-party cookies to track user behavior across first-party sites. Users may use browser settings to block the use of third-party cookies, although "Flash cookies" can be used to potentially respawn such cookies [7].

In the absence of third-party cookies, at least two other approaches could be used to link together the records seen by a third party coming from the same browser. One approach is to simply link together records with the same client IP address, although its use as a unique identifier is uncertain given that the IP address of a home or business machine could change dynamically over time. However, one study [3] found that 72% of clients used a single IP address over a two-week period with higher percentages for shorter durations. A more recent three-week study [6] found 95% of repeat clients retained the same IP address. Thus at a minimum the IP address could be used to increase the confidence in linking records by a third-party.

A new privacy threat—*browser fingerprinting*—where a client can be identified simply by the characteristics of the browser, independent of the client IP address or nature of the request was the focus of [6]. The characteristics are obtained from the HTTP request headers as well from the execution of JavaScript and Flash within a client's browser. The study found that lists of installed browser plugins and fonts have the highest identifiability followed by the User-Agent header. Browser history is another potential threat for fingerprinting—it has been used to de-anonymize social network users [8].

The fingerprinting study [6] identified a new potential privacy issue, but did not examine whether such techniques are being used by third-party aggregators. We examined the data collected during our interactions with each site and looked for the fingerprinting potential, particularly through the inclusion of plugin or font lists in the information sent to third parties. For fonts, we obtained a list of 24 common and uncommon fonts

installed in our test browser using a publicly available test script. We then searched for these fonts in the data we collected, but found no evidence that fonts in our browser are being retrieved and passed to third parties.

We also obtained the list of plugins installed in our browser and looked for them in our data. We found 43 distinct servers (all belonging to Adobe) where a list of plugins was obtained and passed to them via the Request-URI. Figure 10 shows the string of plugins in the Request-URI for a request to a Health category site URL, which actually belongs to Adobe. The request sends a sensitive health search—pancreatic cancer—through the `Referer` and `Cookie` headers, via the Health category site. This same plugin string was passed to Adobe in the example shown in Figure 2 (plugin string was not shown in that figure for brevity), which sent full name and email address of the user. Note that this plugin string is generated and passed by the Adobe script running on a Firefox browser. We observed a similar string being generated and passed on the Chrome and Safari platforms, but did not observe such a string with Internet Explorer. This difference is likely explained by an observation in [6] that IE supports "testable", but not "enumerable" functionality for browser plugins. Thus for three browser platforms using this plugin string as well as other information, such as browser User-Agent and IP address, Adobe is in a position to link together with a high degree of confidence a known user from a Health category site with a sensitive search string that the user is issuing on a popular Health Web site. However, we stress that aggregators, such as Adobe, may have contractual agreements with first-party sites that use them for analytics purposes and forbid them from doing such linkage. Else, what may appear as anonymous information on one site can be linked with identifying information from another site.

GET http://std.o.HEALTH.com/b/ss/...global/...
p=Google Talk Plugin;Google Talk Plugin Video Accelerator;
Adobe Acrobat;Java Deployment Toolkit 6.0.210.7;
QuickTime Plug-in 7.6.6;Mozilla Default Plug-in;
Google Update;Shockwave Flash;Java(TM) Platform SE 6 U21;...
Referer: http://www.HEALTH.com/search/...?query=**pancreatic cancer**...
Cookie: ... s_query=**pancreatic cancer**

Figure 10: Potential Linkage of Sensitive Records Using Browser Plugins

4. EXISTING PRIVACY PROTECTION MEASURES AND NEW PROPOSALS

An earlier examination [14] of current measures to address the privacy problem discussed the role of technology, legislation, and economics, concluding that a combination of these three angles is needed for privacy protection. We next categorize current technologies providing privacy protection and examine their effectiveness in light of our recent leakage disclosures [13, 16] and those in Section 3. We also examine the impact of these technologies on the ability to link together information using other mechanisms as described in Section 3.2.

Two new proposals for protection have been made recently. One was a simple HTTP header addition to request aggregators and ad networks not to track users [5]. The second was a comprehensive report [10] on protecting consumer privacy released by the U.S. Federal Trade Commission in December 2010.

The potential contribution to privacy protection by first-party sites has been largely ignored. Users seek content on the Internet by communicating directly with the first-party sites who have a strong economic incentive in maintaining their viewership count and brand value. Beyond the largely-opaque privacy policies, explicit support for privacy protection by first-party sites might increase their sense of trust with users.

We examine each of these three issues in this section.

4.1 Privacy-Related Work

We now enumerate existing privacy protection measures (shown in **boldfont** for subsequent reference) available to users and the two new protection proposals.

1. Blocking requests to targeted third parties: This **block** measure includes using an advertisement blocking tool (e.g., AdBlock Plus [1]), or a browser built-in (IE9 [2]), to syntactically block selected third parties via server/domain name. Another measure **block-hidden** [15] determines the true source of hidden third-parties by examining their authoritative DNS servers.

2. Refusing cookies to prevent tracking: Browsers can be set to refuse all cookies (**nocook**) or just third-party cookies (**no3rdcook**).

3. Disabling script execution: JavaScript execution can be disabled (**nojs**) either permanently via the browser or selectively via a tool such as NoScript [17].

4. Filtering protocol headers: This is done via extensions or at an intermediary and includes the **referer** measure available in some browsers to modify or remove the `Referer` header in an HTTP request.

5. Anonymizing the user and user actions: One such **anon** measure is anonymizing user's IP address via an anonymizing proxy or by using Tor (https://www.torproject.org/).

6. Opting out of tracking: This can be done via **opt-out** to evade tracking (via cookies) by an aggregator using tools such as the Firefox TACO extension [21] that sets persistent opt-out cookies. Unfortunately some aggregators continue to track when the cookie is present and just not use the information to serve targeted ads [19].

7. Do-Not-Track HTTP header proposal: Researchers proposed in early 2010 that browsers add a HTTP Do-Not-Track-Header (**DNT-Header**) [5] to allow users to express their interest in not being tracked by any aggregator or ad network. However, the extent to which third parties would honor such a header is unknown.

8. FTC consumer privacy protection proposal: The U.S. Federal Trade Commission released a report [10] in December 2010, aimed at policymakers and the industry—stating that companies do not adequately address consumer privacy concerns and that information (such as privacy policies) and choices (various privacy settings) available to users are confusing. The report was based on three meetings held by the FTC in which privacy activists, researchers, technologists, and aggregator company representatives participated. Noting the potential benefits to users about information flow it pointed out the asymmetricity with respect to the low cost of invisible data collection and potential harm to consumers. Additionally, off-line information is being increasingly linked with on-line tracking data leading to easier identification of users. The report seeks a modus vivendi with users and provides input to possible legislation in the U.S. Congress. The report advocates the Privacy by Design [4] initiative, which seeks pro-active embedding of privacy at design stage, defaults to be set to private, transparency about user's information, and access to all user-related sensitive data stored in aggregators.

4.2 Effectiveness of Protection Measures for Leakage and Linkage of Information

With these existing protection measures and proposals, we use the results in Section 3 and from our earlier work to evaluate the effectiveness of prevention measures to protect against known types of leakage and linkage of information. We discuss three scenarios where information is passed to a third-party: 1) typical Web access scenarios with *expected* information shared; 2) *known* to the research community and a demonstrable vector of privacy leakage that is largely unknown to most users; and 3) *potential* linkage that can be demonstrated, but is not yet confirmed as being used by aggregators. We examine situations for each of these scenarios and identify measures that prevent such leakage and linkage. Table 2 summarizes the results.

The first set of scenarios shown in Table 2a are common in the Web and regarded as expected behavior. They provide non-user specific data (such as IP address and browser information) to third parties that could later link the data with more private information.

User visit: Knowledge of a user's visit to a first-party site by a third-party present on the first-party (e.g. ad.doubleclick.net) page—only prevented by **block**.

Hidden third party: Knowledge of a user's first-party visit site by a hidden third-party server—prevented by **block-hidden** to defeat the hidden DNS delegation.

3rd-party tracking linkage: Linkage of received records via tracking by a third-party through third-party cookies. This type of tracking is typically associated with behavioral advertisement and is prevented by **block**, **nocook**, **no3rdcook** and **opt-out**.

1st-party tracking linkage: Records are linked by third-party JavaScript execution that uses first party cookies to store the resulting information. Since users are less likely to block first party cookies, such tracking is prevalent on the Web today. The analytics code is executed by the third party as an add-on to the first party. However, there is a risk that data generated using this mechanism could be used by the aggregator. The use of **nocook** or **nojs** can prevent such potential leakage.

The next set of scenarios are forms of private information leakage to a third-party documented in research literature; these are shown in Table 2b.

Leakage via Referer: User's personal/sensitive information leaks via the HTTP `Referer` header to a third-party server as seen in Figures 1, 5, 6, and 7. This is blocked by **block** and **referer**.

Leakage via cookies: Private information leakage to third-party via first-party cookies, as shown in Figure 2, can only be reliably prevented by **nocook** and **block-hidden**.

Leakage via JS: Leakage of account information to a third-party due to JavaScript execution, seen in Figures 3 and 4—prevented by **block** and **nojs**.

The final set of scenarios, shown in Table 2c, deal with third parties ability to potentially link together information. We can demonstrate the potential, but not confirm such linkage is in use.

Linkage via IP addr: Linkage of records using the IP address of client—prevented by **block** and **anon**.

Linkage via Flash cookies: Linkage of records using Flash cookies [7] to track behavior and potentially respawn deleted traditional cookies—prevented by **block**.

Linkage via fingerprint: Linkage of records using the set of plug-ins loaded into a browser as a potential means of browser fingerprinting, such as shown in Figure 10—can be prevented by **nojs** and **block** (if the third-party can be identified by its domain name).

Linkage via GUID: Linkage of records using globally unique ids (e.g. email address, OSN identifier), such as shown in Figure 9, which can be prevented by **block** (if a known third-party domain) and **referer** (if data is leaked via the `Referer` header).

Linkage w/ Other Sources: Linkage of information using other sources such as search strings, email and public information [22]. This type of linkage is *not* prevented by any of the measures.

4.3 Shortcomings of Privacy Protection Measures and proposals

Most privacy protection measures are *not* effective in preventing many types of leakage and linkage. The technique that provides protection in most scenarios is, oddly enough, an advertisement blocker (Table 2). This

Table 2: Effectiveness of Protection Measures for a) Expected; b) Known; and c) Potential Leakage and Linkage Scenarios

		Protection Measure							
	Leakage/Linkage Scenario	block	block-hidden	nocook	no3rdcook	nojs	referer	anon	opt-out
a) Expected	User visit	X							
	Hidden third party		X						
	3rd-party tracking linkage	X		X	X				X
	1st-party tracking linkage			X		X			
b) Known	Leakage via Referer	X					X		
	Leakage via cookies		X	X					
	Leakage via JS	X				X			
c) Potential	Linkage via IP addr	X						X	
	Linkage via Flash cookies	X							
	Linkage via fingerprint	X				X			
	Linkage via GUID	X					X		
	Linkage w/ Other Sources								

result is particularly notable in light of a recent report that similar protection is to be built-in for the planned new versions of the most popular browser: IE9 (Microsoft's Internet Explorer [2]). While requiring maintenance and agreement on the appropriate blocking set and effective when third parties can be identified based upon their domain name, this technique is not reliable in protecting against cases where a hidden third-party server is used. New measures such as the technique to block based on the true source of such hidden third parties are needed. Blockage of requests can also create usability concerns for the rendered content.

As shown in Table 2, refusal of cookies or disabling JavaScript provides protection in some situations, but as we found in [12], these actions can have negative usability consequences in terms of sites not working properly or pages not rendering properly. The remaining currently available protective measures each have limited usefulness in preventing leakage and linkage.

In comparison to a persistent opt-out cookie, it is argued that the Do-Not-Track header proposal sends a clear signal that a user does not want to be tracked [19], but its success really depends on third parties honoring it. Without enforcement by government agencies or technology that demonstrates lack of such tracking, the effectiveness of this proposal against leakage and linkage is unknown. Additionally, there is no protection for the leakage of information via first-party sites.

The FTC consumer privacy protection report does not adequately address numerous privacy concerns: (1) examination on whether there are adequate safeguards on linkage of existing data when economic acquisitions of aggregators occur (an increasingly frequent phenomenon); (2) verification that aggregators adopting protection measures are indeed abiding by its provisions along with enforcement for violators; (3) visibility of user data obtained and linked together by aggregators along with means for users to view and delete data stored about them by an aggregator—the "right to be forgotten" [18] ensuring that users can erase personal data so it is not retained over time; and (4) the potential tracking of users by other types of third parties, such as those hidden via DNS delegation or even Content Distribution Networks, which have a large presence on the Web.

4.4 Role of First-Party Sites

A key failure of the FTC report is largely ignoring the role of first-party sites in safeguarding consumer privacy. The measures shown in Table 2 are protective measures that *users* could take. However, first-party sites (such as OSNs) should ensure that they take additional steps to protect users from leakage. Even if the leakage is enabled via external applications, the OSNs should be held responsible for enforcement of acceptable use policies. Transparency by first-party sites should be mandated: any legal contracts with aggregators on the latter's ability to obtain data should be disclosed similar to privacy policies. If first-party leakage of data to third parties is inadvertent then sites should be re-engineered to prevent it. Sites (e.g., mobile OSNs) that encourage linking of actions across other sites should constantly disclose the reach of the user's actions so that users are aware of the differences in privacy settings across sites and the possible leakage as a result.

First-party sites should play a *custodial* role in protecting the privacy of users and help prevent any possible leakage. This includes both *explicit* leakage (users wishing to block such leakage have to do it for every server contacted as a result of their action) and *implicit* leakage (information is primarily leaked via `Referer` or `Cookie` headers and can be blocked by single browser settings). Explicit leakages including private information is leaked in a Request-URI or in the body of a POST request, is entirely under the ambit of first-party sites and they should be mandated to block such leakage. Implicit leakage can be a shared responsibility between users and first-party sites with the first-party site warning the users. However, any information that first-

party sites may be populating in a browser header needs to be cleansed of user's private information that is in its custody. Specifically all examples of leakage in Table 2b could be prevented by actions taken by a first-party site.

For example, leakage via JavaScript execution can be prevented by first parties not making private information available to third-party code. Leakage of information via the `Referer` header can be prevented by not passing user information via the Request-URI, but by using a HTTP POST method and passing the information as part of the body of the request. Similarly, first-party sites can prevent leakage to hidden third-party servers either by not using such servers or alternately changing how cookies are set for a first-party domain.

5. SUMMARY AND FUTURE WORK

We have studied hitherto unexamined privacy leakage in a broad spectrum of Web sites where users establish identities and provide personal information. 56% of the 120 popular sites in our study (75% if we include userids) *directly* leak sensitive and identifiable information to third-party aggregators. In new vectors of linkage, we observe cases where globally unique identifiers, such as site userids and email addresses, as well as browser fingerprinting with plugins can be used to link records. Sharing of information by a first-party site with third parties that appears legitimate and even covered by a privacy policy can lead to unhappy surprises when it is linked to an identity via another site. It should be noted that there may be private contractual agreements between aggregators and first-party sites that forbid aggregators from linking information they may receive as a result of user's interaction.

We show the limitations of existing measures and the significant shortcomings of newer proposals in protecting against vectors of privacy leakage and linkage. The growing disconnect between the protection measures and increasing leakage and linkage suggests that we need to move beyond the losing battle with aggregators and examine what roles first-party sites can play in protecting the privacy of their users.

6. REFERENCES

[1] AdBlock Plus. http://adblockplus.org.

[2] Ed Bott. IE9 and tracking protection: Microsoft disrupts the online ad business, February 13, 2011. http://www.zdnet.com/blog/bott/ie9-and-tracking-protection-microsoft-disrupts-the-online-ad-business/3004.

[3] M. Casado and M. Freedman. Peering through the shroud: The effect of edge opacity on IP-based client identification. In *NSDI*, April 2007.

[4] Ann Cavoukian. Privacy by design, 2010. Information & Privacy Commissioner, Ontario, Canada. http://privacybydesign.ca/.

[5] Do not track: Universal web tracking opt-out. http://donotrack.us/.

[6] Peter Eckersley. How unique is your web browser? In M. Atallah and N. Hopper, editors, *Privacy Enhancing Technologies*, volume 6205 of *Lecture Notes in Computer Science*, pages 1–18. 2010.

[7] Soltani et al. Flash cookies and privacy, August 10 2009. http://ssrn.com/abstract=1446862.

[8] Wondracek et al. A practical attack to de-anonymize social network users. In *IEEE Symposium on Security and Privacy*, May 2010.

[9] Fiddler web debugging proxy. http://www.fiddler2.com/fiddler2/.

[10] Protecting consumer privacy in an era of rapid change, Dec 2010. Federal Trade Commission. http://ftc.gov/os/2010/12/101201privacyreport.pdf.

[11] Robin Good. Online advertising management: Ad network defaulting and daisy-chaining for ad revenue optimization. http://www.masternewmedia.org/online-advertising-management-ad-network-defaulting-and-daisy-chaining-for-ad-revenue-optimization/.

[12] B. Krishnamurthy, D. Malandrino, and C. Wills. Measuring privacy loss and the impact of privacy protection in web browsing. In *SOUPS*, 2007. http://www.research.att.com/~bala/papers/soups07.pdf.

[13] B. Krishnamurthy and C. Wills. On the leakage of personally identifiable information via online social networks. In *WOSN*, 2009.

[14] Balachander Krishnamurthy. I know what you will do next summer. *ACM SIGCOMM CCR*, 40(5), 2010.

[15] Balachander Krishnamurthy and Craig E. Wills. Privacy diffusion on the web: A longitudinal perspective. In *WWW*, 2009. http://www.research.att.com/~bala/papers/www09.pdf.

[16] Balachander Krishnamurthy and Craig E. Wills. Privacy leakage in mobile online social networks. In *Proceedings of the Workshop on Online Social Networks*, Boston, MA USA, June 2010. USENIX.

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

[18] Leigh Phillips. EU to force social network sites to enhance privacy. *The Guardian*, March 16 2011. `http://www.guardian.co.uk/media/2011/mar/16/eu-social-network-sites-privacy`.

[19] Rainey Reitman. Mozilla leads the way on do not track, January 24 2011. https://www.eff.org/deeplinks/2011/01/mozilla-leads-the-way-on-do-not-track.

[20] Emily Steel. A web pioneer profiles users by name. *Wall Street Journal*, October 25 2010. http://online.wsj.com/article/SB10001424052702304410504575560243259416072.html.

[21] Targeted Advertising Cookie Opt-Out. http://taco.dubfire.net/.

[22] Pete Warden. What can i find out about you if i know your email address?, December 2009. http://petewarden.typepad.com/searchbrowser/2009/12/what-can-i-find-out-about-you-if-i-know-your-email-address.html.