Study on the Leakage of Private User Information Via a Range of Popular Websites

A Master's Thesis

by

Konstantin Naryshkin

December, 2010

Advisor: Prof. Craig Wills

Reader: Prof. Kathryn Fisler

Department of Computer Science
Worcester Polytechnic Institute
100 Institute Road
Worcester, MA 01609

Abstract

On the modern web, many sites have third party content, be it through maps, embedded objects, ads, or through other types. Users pay little attention to the source of this content since it is such a common occurrence. Unfortunately, this content can be an avenue for third parties to discover private information about the user. Previous work has found these types of leaks in social networking sites. By logging headers during the usage of 120 sites across 12 major categories, we were able to find leakage of a user's private information occurring on many other types of popular web sites. We found leakage on 75% of the sites we looked at and at least one instance in each of the categories. Based on the leaks we found, we propose a classification of the types of leakage that can occur via the HTTP header and use this system to analyze our results.

Acknowledgments

First of all, I would like to thank the people who made my thesis possible. The largest part of the credit has to go to my advisor, Prof. Craig Wills, for helping me accomplish my goals for this project and keeping me on schedule. Without him I would have been utterly lost and would not be able to accomplish half of what I did. I would also like to thank my reader, Prof. Kathryn Fisler. Without her, this report would not be possible and all my work would be meaningless since I would be stuck in grad school for ever. Finally, I would like to thank Dr. Balachander Krishnamurthy for providing feedback on this work, acting as an outside view on the methodology, and co-authoring a paper which will bring these results to the masses.

I would also like to thank everyone who I have met at WPI over the last four and a half years. I would like to thank the professors who taught me the classes I will need to succeed in the real world and tolerating the occasional late homework. I would like thank the TAs for providing comprehensible explanations of the professors' lessons and grading the above mentioned homeworks. I would like to thank my classmates past and present. They provided me with countless hours of amusement and made sitting in a lab and working on homework assignments palatable. I would like to specifically thank the classmates with whom I have had the fortune of doing group work. They provided the best of times and the worst of times.

Finally, I would like to thank my friends and family for tolerating not seeing me for weeks as worked on the various projects that I required for the completion of my degrees, especially as I worked to finish this thesis.

Table of Contents

Table of Figures

Table of Tables

1 Introduction

Due to the interconnected nature of the World Wide Web, it is easy for a site to access content from any number of third-party sites. It is done through embedded objects such as videos and maps, hotlinked images, ads, external JavaScript imports, or any number of other ways. In all of these cases, a user's browser loads content from a third-party[1] site without any explicit action on the part of the user. Each time the content is loaded, it presents an opportunity for a user's private information such as name, email address, hometown, or buying habits to leak to the third party hosting it. Due to the ability of a website to repost this information to a trusted site via an API such as Facebook Connect, there is another avenue for information to leak. This threat is increased even further when using smart phones due to the ability of modern phones to provide the exact location of a user to an application.

We had previously looked at the work of Krishnamurthy, Wills, et al [1,2,3] as well as assisted in data gathering for [3]. Their work looked at the leakage of private information to third parties on social networking sites. Based on their results and our contributions to their data, we saw an opportunity to apply a similar methodology to look at other categories of sites and compare the results from those sites to social networks. Though none of the categories that we looked at stored as much information as social networking sites, it is still possible for the

[1] We use 'third-party' to refer to a site, other than the one that the user is looking at, that provides content that is displayed on the page. We use the domain name to decide if content is third-party. Content coming from the same host as the page the user is looking at is 'first-party'. Content not hosted on the same domain is third-party.

information that these sites do have to leak. Because of the degree to which leakage occurred on social networks, we conjectured that we would be able to find similar types of leakage across other categories of sites.

In this thesis, we explain our study as well as look at our results. We begin with Chapter 2 looking at the context of our problem. In particular, we focus on defining what is private information, what constitutes leakage of it, and what solutions exist to prevent such leakage. We then discuss our methodology in Chapter 3. We look at selecting categories and sites to look at. We also look at how we gathered data for each site. We continue with Chapter 4 containing the Analysis of Leaks. In this chapter we cover the types of leaks that we found, the frequency of leakage in each of the categories, and the recipients of data that is leaked. Our last paper chapter, Chapter 5, is Conclusions. In it, we evaluate our hypothesis and look at future work. Finally we include Appendices containing a full list of sites that we looked at and a glossary of terms that we use in this paper for the benefit of the reader.

2 Background

2.1 What is Private Information

In order to study leakage of private information, we first need to decide what is **private information**. We define it as any personal information about the user as well as any information that the user gives to a site that the user may not like to be widely known. To define personal information, we turn to literature. A directive by the European Parliament [4] defines it as "information relating to an identified or identifiable natural person". The National Institute of Standards and Technology, an agency of the US government, defines personally identifiable information "any information about an individual…, including (1) any information that can be used to distinguish or trace an individual's identity … and (2) any other information that is linked or linkable to an individual" [5]. In both cases, personal information is defined as including any information that can be associated with a given user. This information should not be limited to data that uniquely identifies an individual (referred to as *identified data* [6]). There are studies [7] showing that several pieces of data that do not separately identify a user can be combined to do so. This type of data is referred to as *identifiable data* [6]. Information that we felt a user would not want widely known includes medical information that the user was interested in, travel arrangements that the user was making, and the hobbies of the user.

We define a **private information bit** as a single datum of private information. It is a fact that a user can disclose to one or more websites and it is the smallest amount of information that could be known. Examples include a birth year or zip code.

2.2 Leakage of Private Information

We define **leakage of private information** as any situation in which a user's private information is disclosed to a third party without a user explicitly requesting such disclosure. Leakage is a well-known occurrence [6]. Despite the negative attitude about it, in many instances it is considered an unwanted, but tolerated menace. In the US, both the NIST [5] and Office of Management and Budget [8] have warned government agencies about the dangers of breaches of personally identifiable information. In 1995, the European Union has defined the privacy of one's personal information as being a fundamental human right [4]; it reiterated and strengthened their stance in 2010[9].

2.2.1 User Opinions of Leakage

Over the years, there have been numerous studies that focused on user opinions about companies harvesting user information. More than 80% of Americans report having a negative opinion of corporations collecting private information [10] and 82% of Europeans have little trust in the Web with private information [11,12]. Studies found that, if given the option, users will pay a relatively small premium on an item if they feel that they are buying from a more secure retailer [13]. Several studies implied that users found the practice of harvesting user information and storing it in a database for future marketing purposes to be unethical [14,15]. More than 90% of Europeans seem to agree, since they do not approve of the usage of their private information for targeted ads [11]. The most shocking studies are those discovering that users are not aware of the full extent of leakage of their private data [16,17].

2.3 Threat Posed by Leakage on the Web

Subsection 2.2 establishes that the leakage of private information is generally regarded as a negative event. The next logical step is to look at how the Web can act as an avenue for a user's private information to be gathered and leaked. As noted in [7], it is possible to collect bits of private information from various sources to build a more complete profile of the user. One of the common ways to unite multiple bits is to use a persistent cookie in the browser as a user identifier [18]. The frequency with which it occurs has prompted the World Wide Web Consortium to add a note on how to inform users about it in their specification for machine readable privacy policies[19]. Another way to gather private information about a user is by inferring it from user history as demonstrated by [20]. Third parties can gather a user's history either by tracking the user's cookie or by using a well-known CSS exploit [21] that is not likely to be fixed in the foreseeable future [22]. Location bits about the user can be inferred from the IP address [20,3]. This research shows yet another way that bits of private information can get into the hands of third parties. All of these bits can be combined by a third party to form a profile of the user.

Once all of these bits of private information are gathered, they can be used for purposes such as targeted advertising. The practice of targeted advertising involves using a user profile to display ads that are believed to appeal to the interests of the user. The practice is assumed to be widely used by advertisers [23,24]. It has also been shown that the ads shown to a user in targeting advertising may expose information about the user to the advertised company [24].

2.4 Preventive Measures

For all of the threats to private information, there exist countermeasures. Unfortunately, none of the solutions are perfect. For all of them, the biggest issue is that users do not use them [25].

2.4.1 Cookie management

If users are concerned about being tracked using cookies, they can manually delete cookies regularly or choose not to accept them at all. Unfortunately, taking any of these actions can degrade the user experience as noted in [26]. It is also not effective in every case. The users can still be tracked by using alternative cookie implementations through technologies like Flash local shared objects [18]. The users can also be tracked using browser fingerprinting as demonstrated by [27,20] or simply though the IP address.

2.4.2 Network Anonymizers

Another preventive measure is to anonymize the user's connection through technologies such as Tor [28] and a HTTP proxy front end like Privoxy[2]. The weakness of this method is that it is protocol agnostic. Being protocol agnostic means that it only stops leakage at the network layer, but does not stop any leakage at the HTTP level, which includes most of the leakage that we mentioned. It stops IP address leakage, which can be used to infer the user's location or track a session.

[2] http://www.privoxy.org/

2.4.3 Browser Plugins

There exist a number of browser plugins that are capable of blocking advertisements and filtering JavaScript. Popular choices include Adblock Plus[3] and NoScript[4]. Like disabling cookies, these plugins can hurt the user experience [26]. Since they stop third-party ads from loading, they likely interfere with a sites ability to sell advertising space. This interference has ethical implications, since the user is getting the website contents without the site owner being paid [29]. It also falls into a legal gray area, since the user may be making an unauthorized modification to the intellectual property of the site's author [29,30]. There exists a theoretical solution proposed by [31] that involves the user running an ad server locally. This solution eliminates any leakage since none of the user's information leaves the local machine. The system requires a complicated infrastructure to distribute the ads to all the local servers, verify the authenticity of ad responses, and to make sure that a user's actions are truly kept secure.

2.5 Summary

We define private information as a user's personal information as well as any other information that a user does not want to be widely known. From literature, we chose a definition of personal information that included both information that identifies a user, and information that can be linked to a user. Again, to match literature, we chose to define leakage of personal

[3] http://adblockplus.org

[4] http://noscript.net/

information to be any instance of a user’s private information being revealed to a third party without the consent of the user. Based on this definition, we found evidence of both legal and popular support for the limiting of such leakage. We also showed how instances of leakage can occur during usage of the World Wide Web and how multiple instances can be combined to get a more complete profile of the user. Finally, we discussed several popular solutions for preventing leakage, but noted that all of them had significant drawbacks. In the following chapters, we demonstrate another avenue for leakage to occur on the Web.

3 Study Design

The intent of our study was to look at how prominent privacy leakage was on a number of popular categories of sites. Looking at a variety of sites would allow us to compare the different categories to see if visiting certain types of sites puts a user in greater risk of privacy leakage. We could also look at the overall privacy leakage on the World Wide Web (by examining what portion of the categories show potential for leakage). Once we had selected a set of categories and the sites within each category that we would be looking at, we designed a uniform way to look at the type of information that would be leaked about the user when the site was used.

3.1 Selecting Categories and Sites

One of the first issues that we encountered with the project was to find an objective way to select a set of categories and the sites within them while still having a set of sites that were of research interest. The first step to accomplishing this task was to set up several criteria that every site that we would look at would need to fulfill. The first criterion was that the site must collect enough information about a user so that it would have something to leak. We insured this criterion was satisfied by only looking at sites that allowed a user to create an account on the site and making sure that the site offers additional functionality exclusive to users who created accounts. Another criterion was that the sites must have a large enough population of users about whom they have data. To verify this criterion, we tried to find information about the number of registered users on the site and only look at sited with at least 100,000 registered users (although for most of the sites that we decided to include, the number was in the millions). In general, the number of users on a site was a self-reported statistic. We were not able to find this information about all the sites, but those that we allowed were popular sites in a category where other sites had significant populations. Our third criterion was that each site must offer functionality and features consistent with other sites in its category. For categories in which we found more than one popular set of functionality, we selected the set of sites which included the most popular ones. Our final criterion was that the site could be studied without disclosing too much about who we were. This criterion excluded sites that required a method of payment to create an account or ones that verified identity beyond an email address.

We based our categories on the categories and sub-categories used by Alexa[5], a site popularity ranking. We filtered the categories and sub-categories to find ones that had 10 sites fulfilling our criteria near the top of the listing. Categories in which we could not find ten sites matching our criteria among the top sites were dropped. Using Alexa insured that the sites in each category were the most popular ones that we could look at. Using this procedure, we came up with a list of ten categories that we wanted to look at: Arts, Employment, Video Game News and Reviews, Photo Sharing, News, Travel, Shopping, Relationships, Generations and Age Groups, and Sports.

In addition to these ten categories, we included two more in our study. One was Online Social Networks (OSN) since they were look at previously in [3] and since the sites have huge numbers of registered. We used the same criteria for site selection as the other categories and

[5] http:// www.alexa.com

gathered fresh usage for all the OSN sites. The other category that we decided to add was Health. We wanted to include this category because of the sensitivity of the information that a user provides to a health site. This information is of a private nature and if it were to be leaked to third parties, it could be combined with other information about the user and hurt the user in pursuits such as finding employment or purchasing insurance. To allow for the inclusion of this category, we needed to loosen the requirement of sites allowing users to register. Some the popular sites in this category do not allow user registration but still have plenty of private information about the user because the sites have other private information, such as search terms, that are of interest to third parties. We created an account on these sites as well if doing so was possible.

3.2 Gathering Site Interactions

Once we had established a list of sites, we worked to create a log of typical interactions between the site and a user. The approach that we used is the same one that was used by [2]. We set up a browser to route all of its traffic through the Fiddler Web proxy[6]. There was a set of actions that was common to almost all sites that we made sure to do. We started by creating an account. We would then confirm any verification emails, as appropriate. We would try to create and edit our user profile if the site provided one. We would then explore the capabilities offered by the site to registered users. The actions we looked at often included browsing the contents of

[6] http://www.fiddler2.com/fiddler2/

the site, using any search capabilities that the site offered, and posting comments on and reviews of the site contents. Many sites offered a “remember me” functionality. When one was present, we selected it, as many sites implement the feature by storing private information in a cookie (which can be leaked to a third party). There were many other capabilities that varied by category that we tried:

- Arts: While accessing videos or music we used play/pause/resume when appropriate. We also posted comments about content and personalized the site with favorites as only registered users are allowed to do so.
- Employment: Uploaded resumes, searched and applied for jobs, and signed up for email alerts.
- Video Game News and Reviews: Registered users alone can review and comment on games, add games to a list of owned games, and join groups. Beyond these actions we read and contributed to forums available for users to interact with each other.
- Photo Sharing: Created a photo gallery, uploaded and commented on images, shared images with guests. We also participated in user communities if present.
- News: Took advantage of registered users’ capabilities such as commenting on articles and marking articles as favorites.
- Travel: Read recommendations and information about destinations, booked travel arrangements (without actual payment), and saved flight itineraries.
- Shopping: Added items to a shopping cart (no actual purchases done), looked for a “nearest store” for sites with physical stores, and created a wish list when possible.

- Relationships: Took compatibility quiz, browsed local area members, and tried to contact examined matches.
- Generations and Age Groups: Joined groups/forums and posted comments, participated in contests for registered users.
- Sports: Looked at sports scores, read articles, and viewed video clips. As a registered user we commented on articles, checked on “favorite” teams, and participated in fantasy sports contests.
- OSN: Used results obtained as part of [3], which included viewing user profiles, searching for friends, posting messages, uploading photos and installing/running external OSN applications.
- Health: Searched for and browsed health conditions, participated in surveys and took quizzes. As a registered user we used community features and posted comments.

There were some additional actions that, though not common to any category of sites, we encountered occasionally and took advantage of when we did. When a site had a forum, we read topics and contributed to the discussion. If a site had the ability to link a user’s account on this site to an account on another site, we took advantage of it. This feature creates the potential that the login information from the other site is stored by this one and subsequently leaked. Some sites offered the ability to share an article or post from this site on another one; we tried this action for similar reasons.

Once we had session information from all of the sites, we looked through all of the headers as well as POST data for user ids, user names, and pieces of private information that were getting sent to third parties. These headers included bits such as full name, zip code, and email address. For the most part, these searches could be done in an automated fashion, but we did need to eliminate false positives by hand and some bits, such as gender, had to be searched for through manual inspection. This process gave us a listing of all leaks to third parties as well as how the information was leaked and all the recipients of it.

There are a number of reasons for us to not be able to observe a leak. Though we search for common strings in our logs, there is a chance that the string was passed in such a way as to not match our regular expressions. We did not have a way to decode encrypted transmission such as SSL. We did not have a way to read data that was encrypted by a script before it was stored or transmitted. It is important to note that we could only analyze the leaks that we observed. As such, our results are the **lower bound** of information that leaked.

3.3 Summary

To measure privacy leakage on the World Wide Web, we decided to look at 12 popular categories of sites. From each category, we selected the 10 most popular sites that gather and store user information. On each of the sites, we created a user account and carried out the types of actions that would be done by users on that site. All of our interactions were logged by a proxy. We then inspected those logs to see if we could find any evidence of users' private information being leaked. Our findings are discussed in the next chapter.

4 Analysis of Leaks

There were several types of information that we were looking for in the site data that we collected. One was a way to classify the type of leakage that we observe. Another was to look at how prominent leakage was in each of the categories that we looked at. A final one was to look at the recipients of the data.

4.1 Types of Leakage

The most obvious types of leakage that we saw occurring was what we refer to as **direct leakage**. This type of leakage is leakage that results in one or more bits of a user's private information being transferred in plain text form to a third party as part of an HTTP header. There are a number of ways that this leakage can occur. The six that we have found in the sites that we looked at are: leakage via cookies, leakage via request URL, leakage via referer, leakage via title, leakage during signup, and leakage of search terms.

4.1.1 Leakage via Cookies

Leakage via cookies occurs when the user's cookies are sent to a site other than the one that set them. Many of the sites that we looked at store some private bits in the cookies, so when a cookie are leaked, any information stored in it is leaked as well. Cookies generally leak because some metrics companies, most notably Omniture, have a sub-domain of the sites main domain redirect to one of their servers. Borrowing the term from [32] we refer to these sites as **hidden third parties**. The user's browser, assuming that the sub-domain is the same site as the main site, sends any cookies associated with the domain to the sub-domain. An example of this leakage is in Figure 1 (note that the private information being leaked is shown in a **bold font** in all the leakage examples). In it we see that aarp.org, one of the sites in the Age Groups category, stores the user’s email address, full name, and zip in a cookie sent to metrics.aarp.org. A DNS lookup reveals that the sub-domain points to 207.net (which is one of Omniture’s domains).

GET: http://metrics.aarp.org/b/ss/aarpglobal/...

Host: metrics.aarp.org

Referrer: http://www.aarp.org/

Cookie: ...a=jdoe&e=**jdoe@email.com**&f=**John**
&l=**Doe**&...&p=**12201**...

Figure 1 Leakage of Email Address, Full Name, and Zip Code to a Hidden Third Party

4.1.2 Leakage via Request URL

Leakage via request URL occurs when the user loads a third-party site URL that contains some bits of the user’s private information. This type of leakage generally occurs because the site that the user is visiting puts information into the URLs of scripts, images, or multimedia embedded within a page. This type of leakage can be seen when the information is an argument passed via GET to a third-party script. An example of this type leakage is presented in Figure 2. In it, we see that webshots.com, one of the sites in the Photo Sharing category, passes the user’s age and gender to doubleclick.net.

GET: http://ad.doubleclick.net/ … **a=25;g=M**;dcopt=ist;sz=300x250;ord=72457?

Host: ad.doubleclick.net

Referrer: http://community.webshots.com/myphotos? ...

Figure 2 Leakage of Gender and Age via a Request

4.1.3 Leakage via Referer

Leakage via referer occurs when the sites that the user is looking at contains bits of private information in the URL. This information could be there because the information is being passed by the site via a GET request, because the site got the information due to leakage via request URL, or because the URL just happens to contain a bit of private information in it. Any pages linked to by the current page as well as any images, media, or scripts embedded in the current page get the pages URL in the referer field when loaded. Figure 3 is an example of this type of leakage. We see guardian.co.uk leaking the user's email address to scorecardresearch.com. The leakage occurs because both bits of private information occur in the URL of the page that the user is looking at.

```
GET: http://b.scorecardresearch.com/r?c2=6035250&d.c=gif…
Host: b.scorecardresearch.com
Referrer: http://users.guardian.co.uk/mydetails?
          AU_LOGIN_ID=jdoe%40email%2ecom&AU_PASSWORD=%2d%2
          d%2d%2d%2d%2d&AU_PASSWORD_HASH=d0c9811d9f103499080
          041d756840fdb…
```

Figure 3 Leakage of Email Address via Referer

4.1.4 Leakage via Title

Leakage via title is a special case of other types of leakage. It is leakage that occurs when one or more bits of private information are displayed in the title of a page. Some third-party scripts record the title of the page that is calling them. These scripts end up recording the private information in the title. In Figure 4 we see an example. Hulu, one of the sites in the Arts category, puts the user's full name in the title of a page. JavaScript code running on the page takes the title and passes it as an argument to another script via request URL. Note that the page is also leaking the user's Hulu user id via the referer.

```
GET http://beacon.scorecardresearch.com/... c8=Hulu - John Doe's profile...
Host: beacon.scorecardresearch.com
Referer: http://www.hulu.com/profile/public/|123456789...
Cookie: UID=7c91325-215.167.49.28-1325396156
```

Figure 4 Leakage of the User's Full Name via Title

4.1.5 Leakage during Signup

Leakage during signup is another special case of other types of leakage (generally leakage via request URL or leakage via referer). It is of interest because it is leakage specific to actions performed during signup. Due to the information that the user reveals to a site during account creation and setting up a profile, it has a lot of potential to leak a large amount of sensitive user information. Leakage during signup also includes leaks of verification email contents by a site, which may contain user name, email, or password. Figure 5 is an example of this type of leakage. When the user creates an account with criticinfo.com, which is part of the Sports category, the user is sent a confirmation email. When the user attempts to confirm the account, the user's email is included in the URL of the confirmation page and is leaked to doubleclick.net via referer.

```
GET http://ad.doubleclick.net/adj/...
Host: ad.doubleclick.net
Referer: http://submit.cricinfo.com/member mgmt/.../confirm.html? email=jdoe@email.com
```

Figure 5 Leakage of Email during Signup

4.1.6 Leakage of Search Terms

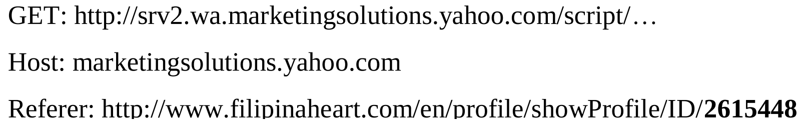
Leakage of search terms is one more special case of other types of leakage. It is leakage, generally via referer, of a user's search terms when a user performs a search on a site. It is noteworthy because of the data that it is leaking. Most other leakage that we saw was leaking information stored in the profile of the user. Search terms may contain information that a user would not normally want to list in his or her profile. We see an instance of this type of leakage in Figure 6. In the figure the user's search for the term "pancreatic cancer" on menshealth.com, one of the Health sites at which we looked, is being leaked to a third party. Users may not want these types of sensitive search terms associated with them.

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

Figure 6 Leakage of a Health Issue as a Search Term

4.1.7 Indirect Leakage

Not all leakage that we observed fit our definition of direct leakage. This other type of leakage was classified as being indirect. **Indirect leakage** occurs when information is passed to a third party that is not in itself private information, but can be used to access private information. This category consisted of information that could be used to access the profile of the user to extract more information. The information that this type of leakage can potentially reveal may be lessened significantly by a user setting aggressive privacy settings on sites. By the same token, an overly trusting user can potentially leak a large amount of information this way. The leakage in Figure 7 is of this type. Filipinaheart.com, one of the sites in the Relationships category, leaks a user’s user id which can be used to access the user’s profile. From that profile, the third party can get access to a user’s city of residence, age, a photo of the user, and details about the user’s sexual orientation, among other bits of information.



```
GET: http://srv2.wa.marketingsolutions.yahoo.com/script/…
Host: marketingsolutions.yahoo.com
Referer: http://www.filipinaheart.com/en/profile/showProfile/ID/2615448
```

Figure 7 Leakage of a User’s Profile That Allows Indirect Leakage

4.2 Leakage by Category

Once we had established the leakage types, we looked at the frequency with which each type occurs in our data. Table 1 shows the count of sites, out of the 10 in each category, that have leakage to a third party. The categories have counts for how many sites have direct leakage, how many have indirect leakage, and how many have any leakage, whether it is direct or indirect. All of the categories had instances of both direct and indirect leakage and most categories had some kind of leakage on at least half of the sites we looked at. The sole exception is shopping. This exception exists largely because shopping sites only have user information to expedite checkout. It is also because shopping sites have a source of revenue other than advertising, so they tend to have fewer ads than sites in many of the other categories. Out of the 120 sites in the categories, 90 of them had some sort of leakage and 67 had direct leakage. These numbers strongly support our hypothesis that we would be able to find leakage across many different categories of sites.

Most of the categories that ended up near the top of the list did so largely due to direct leakage. The Health, Travel, and Employment categories all had much more direct leakage than indirect. In contrast, both the Video Game News and Review and the Generations and Age

Groups categories had significantly more indirect leakage than direct leakage. This contrast seems to suggest indirect leakage and direct leakage are independent of one another. The categories that had a majority of sites with indirect leakage (OSNs, Video Game News and Reviews, Arts, Generations and Age Groups, and Relationships) all have profiles as major features. The categories that do not have indirect leakage on most sites (Health, Sports, News, Travel, and Shopping) were all categories where a profile is not a noteworthy feature, if it is even present. This result supports the obvious idea that having a profile with user information in it can cause that user information to be leaked.

Table 1 Total Sites with Leakage in Each Category

Category	Direct Leakage	Indirect Leakage	Any Leakage
Online Social Networks	7	10	10
Relationships	7	6	10
Health	9	4	9
Travel	9	1	9
Employment	8	2	8
Arts	7	7	8
Video Game News and Reviews	2	8	8
Sports	4	4	7
Generations and Age Groups	2	7	7
News	5	3	6
Photo Sharing	4	5	5
Shopping	3	1	3
Total	67	58	90

Since we had further classifications of direct leakage and since direct leakage is more obviously an issue, we looked at the type of direct leakage that occurred on each site. The results are summarized in Table 2. The categories are sorted in order of how many of the sites exhibit direct leakage. The numbers are a count of how many of the sites in a given category showed this type of leakage. The Any column is the same as the Direct Leakage column in the previous table. Again, a majority of categories have leakage on at least half of their sites. Both travel and health are at the top of the list due to leakage of search terms.

Table 2 Types of Direct Leakage in Each Category

Category	Any	Leakage via Cookies	Leakage via Request URL	Leakage During Signup	Leakage via Referer	Leakage via Title	Leakage of Search Terms
Health	9	1	0	0	0	0	9
Travel	9	0	1	0	0	0	9
Employment	8	0	2	0	5	2	4
Online Social Networks	7	0	3	0	0	5	0
Arts	7	0	3	0	1	4	0
Relationships	7	0	3	0	0	2	2
News	5	3	2	0	0	0	0
Photo Sharing	4	1	2	3	1	0	0
Sports	4	2	0	1	1	0	0
Shopping	3	2	0	0	1	0	1
Video Game News and Reviews	2	1	0	0	0	1	0
Generations and Age Groups	2	1	0	0	0	1	0
Total	67	11	16	4	9	13	25

4.3 Bits Leaked

After looking at where leakage is coming from and where it is going, the next result to look at is what bits are getting leaked. Table 3 contains a list of all bits that we saw leaked ordered by how many sites leak the bit. The table shows how many distinct sites that we looked at leaked the bit and how many distinct third parties received the bit from one or more sites. Among the top bits, we see bits that uniquely identify the user, such as full name which was leaked from 20 out of the 120 sites we looked at and was sent to 20 different third parties. We also see bits that are not unique to a given user, such as age, which leaked from 21 sources to 22 destinations. Notable bits include medical and travel related searches. They are only leaked by 9 out of the 10 sites in their respective categories but end up being sent to more destinations than almost every other bit.

Table 3 Bits Leaked and Counts of Sources and Destinations

Bit	**First-Party Sources**	**Third-Party Destinations**
age	21	22
full name	20	20
zip code	18	29
gender	16	20
city	16	19
email address	15	16
medical searches	9	30
travel related searches	9	25
occupation	4	8
home address	3	9
hobbies	3	1
college	2	6
home phone	2	3
employer	1	6

4.4 Leaked Bits' Recipients

Once we establish what categories are leaking data, the next logical question to ask is what sites are receiving these leaks. Table 4 lists the top 10 leak recipients of direct leakage, in order of the number of sites leaking to them. It also includes the number of sites that leak to this third party and the number of bits of private information that are leaked. Overall, the distribution of recipients has a few receivers with a large number of leaks going to them and many receivers that only get information from a few sites. Missing from this table are content delivery networks, such as Akamai. Though they host content on a server external to the site the user is looking at, the content was still generated by the first-party. We also excluded servers that are owned by the same company as the site the user is looking at and used to host content related to the site.

Table 4 Top 10 Third-Party Recipients of Private Bits from Direct Leakage

Third-Party Domain	First-Party Sites Leaking	Bits Received
doubleclick.net	27	13
google-analytics.com	24	11
scorecardresearch.com	18	7
omniture.com	17	8
atdmt.com	12	9
yieldmanager.com	9	11
2mdn.net	9	9
quantserve.com	8	8
collective-media.net	6	8
2o7.net	4	7

DoubleClick, an advertising subsidiary of Google, is the top recipient of direct leakage. It is getting data from 27 separate sites out of our 120. The bits leaked were: city, zip code, hobbies, gender, age, travel related searches, employer, occupation, college, address, full name, medical searches, and email address. DoubleClick received all but one of the bits that we saw leaked by sites. The only bit that it did not receive was the user’s phone number. The second largest recipient was Google Analytics, which got 11 bits of information from 24 sites. The bits were:

city, zip code, gender, age, travel related searches, occupation, address, full name, medical searches, email address, and home phone number. The recipient in third, Score Card Research, got leaks from 18 sites but only received 7 distinct bits of private information. They are: city, zip, address, full name, medical searches, email, and home phone number.

4.5 Summary

Using the results of our study, we created a system of classification for the types of leaks we found. Broadly, we broke apart the leakage into two main types: direct leakage, which has bits of private information sent directly to the third party, and indirect leakage, which has a key sent to the third party that allows the third party gain access to a collection of person information about a user. The direct leakage was further broken into six types covering all the instances we saw. These types were: leakage via cookies, leakage via request URL, leakage via referer, leakage via title, leakage during signup, and leakage of search terms. Overall, we found leakage on 90 out of the 120 sites that we looked at, including instances of both direct and indirect leakage in every category that we looked at. As expected, categories that tended to contain sites with profiles tended to have a lot of indirect leakage, while categories that did not have many profiles had few such instances.

We looked at the 67 sites that exhibited direct leakage in further analysis. When looking at which third parties got the most leaks, we found that a small minority of them received leaks from a large portion of the sites, but the large majority of third parties only received private information from a few different sites. When inspecting the bits that were leaked, we found that both bits that uniquely identify a user and those that do not got leaked with similar frequency.

5 Conclusions and Future Work

Our hypothesis, that we are able to find leakage of private information on websites other than social networks, was confirmed by the results. By simply looking at HTTP headers, we were able to identify instances of information leakage on 75% of the sites that we looked at. In every category that we looked at, there were instances of both direct and indirect leakage. In almost every category, more than half the sites had leakage of some sort. These results are significant, especially since our findings are the lower bound of the leakage. There exists leakage that our methodology misses since we are only able to look at clear text leaks in the HTTP headers.

There exist several ways that the methodology that we used for this study can be applied to a larger set of sites. The most obvious is to see if the same types of leakage that we saw in our categories also exist in other categories of sites. Our categories were chosen mainly to satisfy a requirement of popularity and fairness of selection. There are a number of categories that can be looked at that have interesting privacy implications or may have results that are different from ours.

Another direction that the work can be taken is to follow the idea laid out in [3] and compare mobile sites to their full counterparts. There exist a number of sites on the Internet that have a full version, a mobile internet version, and an app for one or more cell phone operating systems. The three versions can be compared to see if they all exhibit the same types of leakage. It is possible that other types of leakage can be found which come about due to the properties of a mobile platform.

There exists a lot of work that can build upon our findings. One such idea is creating a tool to prevent the types of leakage that we found. Much of the direct leakage that we found could be prevented if the browser was aware of private information about the user. The browser could then look at what data is being sent to third party hosts and strip out any private information that would be leaked. The process is not perfect since not all possible patterns of leaks can be matched by a regular expression, but one could envision a tool with an acceptable error rate. It would get around some of the problems with current comparable tools, since the user would still have access to the same third party contents as an unprotected user, but the user would be protected from the types of leakage that we identified.

Another security measure that could be implemented is identification of hidden third parties. The browser would then be able to treat them as third-party servers in regards to cookies and other security measures. To accomplish this task, the browser would need to do a reverse DNS lookup on all IPs that it sends data. It would also need to have some ways to differentiate a true hidden third party from a CDN or utility computing service.

In addition, there exists work that we could not do for various reasons but could expand our work. One task is demonstrating the viability of harvesting information from indirect leakage. To do so, one would need to take the headers to a popular third party domain such as doubleclick or google-analytics and create a script to automatically search the logs for data that looks like a profile identifier. The script would then need to go to the page and see what data it could find. If such script were to be written (and there is no reason to say it cannot be), it would show that a malicious third party can harvest indirect leakage form many sites with minimal effort.

An additional follow up work is to look at how the leakage on a site changes over time. Features are constantly added to site and old features are redesigned. There is potential that the changes close up an avenue for data leakage, but it is also possible that the new feature also adds new leaks. Many sites also have advertising campaigns from any number of third parties. Different campaigns may end up leaking different data and they will most likely be leaking to different third parties. This type of study can also show us, if we look at long enough of a sample, if the amount of leakage is increasing or decreasing.

Bibliography

[1] Balachander Krishnamurthy and Craig E. Wills, "Characterizing privacy in online social networks," in *Proceedings of the Workshop on Online Social Networks in conjunction with ACM SIGCOMM Conference,* Seattle, WA USA, 2008, pp. 37-42.

[2] Balachander Krishnamurthy and Craig E. Wills, "On the leakage of personally identifiable information via online social networks," in *Proceedings of the 2nd ACM workshop on Online social networks*, New York, NY, 2009, pp. 7-12.

[3] Balachander Krishnamurthy and Craig E. Wills, "Privacy Leakage in Mobile Online Social Networks," in *Proceedings of the Workshop on Online Social Networks*, Boston, MA, 2010.

[4] European Parliament and the Council of the European Union, Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, November 23, 1995.

[5] Erika McCallister, Tim Grance, and Karen Scarfone, "Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)," *Recommendations of the National Institute of Standards and Technology*, no. NIST Special Publication 800-122, April 2010.

[6] Balachander Krishnamurthy, "I know what you will do next summer," in *ACM SIGCOMM CCR*, 2010, pp. 40-45.

[7] Bradley Malin, "Betrayed by my shadow: learning data identity via trail matching," *Journal*

of Privacy Technology, June 2005.

[8] Clay Johnson III, "Safeguarding against and responding to the breach of personally identifiable information," Office of Management and Budget, Executive Office of the President, Washington, DC, Memorandum M-07-16, 2007.

[9] "Consultation on the Commission's comprehensive approach on personal data protection in the European Union," Unit C3 – Data protection, Directorate-General Justice, Public consultation COM(2010)609, 2010.

[10] Joel Roberts. (2005, October) CBS News Opinion. [Online]. http://www.cbsnews.com/stories/2005/09/30/opinion/polls/main894733.shtml

[11] EurActiv. (2008, May) Regulator warns of mobile Internet privacy concerns. [Online]. http://www.euractiv.com/en/infosociety/regulator-warns-mobile-internet-privacy-concerns/article-172783

[12] EurActiv. (2008, April) Online privacy a concern for EU citizens. [Online]. http://www.euractiv.com/en/infosociety/online-privacy-concern-eu-citizens/article-171742

[13] J Tsai, S Egelman, L Cranor, and A Acquisti, "The Effect of Online Privacy Information on Purchasing Behavior: An Experimental Study," in *The 6th Workshop on the Economics of Information Security (WEIS)*, 2007.

[14] G Nowak and J Phelps, "Understanding Privacy Concerns: An Assessment of Consumer Information – Related Knowledge and Beliefs," *Journal of Direct Marketing*, vol. 6, no. 4, pp. 28-39, 1992.

[15] M Brown and R Muchira, "Investigating the Relationship between Internet Privacy Concerns and Online Purchase Behavior," *Journal of Electronic Commerce Research*, vol. 5, no. 1, pp. 62-70, 2004.

[16] Knowledge@Wharton. (2008, June) Knowledge@Wharton. [Online]. http://knowledge.wharton.upenn.edu/article.cfm?articleid=1999

[17] J Turow, L Feldman, and K Meltzer, "Open to Exploitation: American Shoppers Online and Offline," University of Pennsylvania, 2005.

[18] Ashkan Soltani, Shannon Canty, Quentin Mayo, Lauren Thomas, and Chris Jay Hoofnagle. (2009, August) Social Science Research Network. [Online]. http://ssrn.com/abstract=1446862

[19] Lorrie Cranor et al. (2006, November) The World Wide Web Consortium. [Online]. http://www.w3.org/TR/P3P11/

[20] Craig E. Wills and Mihajlo Zeljkovic, "A Personalized Approach to Web Privacy—Awareness, Attitudes and Actions," Computer Science Department, Worcester Polytechnic Institute, Worcester, MA, Technical Report WPI-CS-TR-10-07, 2010.

[21] Markus Jakobsson, Tom N. Jagatic, and Sid Stamm. Stop-Phishing.com. [Online]. https://www.indiana.edu/~phishing/browser-recon/

[22] Mozilla Foundation. (2002, May) Bugzilla@Mozilla. [Online]. https://bugzilla.mozilla.org/show_bug.cgi?id=147777

[23] Mike Nolet. (2007, February) Mike On Ads. [Online].

http://www.mikeonads.com/2007/02/28/how-do-behavioral-networks-work/

[24] Saikat Guha, Bin Cheng, and Paul Francis, "Challenges in Measuring Online Advertising Systems," in *Proceedings of IMC*, Melbourne, Australia, 2010.

[25] A Acquisti and J Grossklags, "Losses, gains, and hyperbolic discounting: An experimental approach to information security attitudes and behavior," in *In 2nd Annual Workshop on Economics and Information Security (WEIS '03).*, 2003.

[26] Balachander Krishnamurthy, Craig E. Wills, and Delfina Malandrino, "Measuring privacy loss and the impact of privacy protection in web browsing," in *Proceedings of the Symposium on Usable Privacy and Security*, Pittsburgh, PA USA, 2007, pp. 52-63.

[27] Peter Eckersley, "How Unique Is Your Web Browser?," *Lecture Notes in Computer Science*, vol. 6205/2010, pp. 1-18, 2010.

[28] Roger Dingledine, Nick Mathewson, and Paul Syverson, "Tor: the second-generation onion router," in *Proceedings of the 13th conference on USENIX Security Symposium*, vol. 13, 2004.

[29] Noam Cohen, "Whiting Out the Ads, but at What Cost?," *The New York Times*, September 2007.

[30] Jilian Vallade, "Note: AdBlock Plus and the Legal Implications of Online Commercial-Skipping," *Rutgers Law Review*, vol. Spring, 2009.

[31] Saikat Guha, Alexey Reznichenko, Kevin Tang, Hamed Haddadi, and Paul Francis, "Serving Ads from localhost for Performance, Privacy, and Profit," in *Proceedings of the*

8th Workshop on Hot Topics in Networks (HotNets '09), New York, NY, 2009.

[32] Balachander Krishnamurthy and Craig E. Wills, "Privacy diffusion on the web: A longitudinal perspective," in *Proceedings of the World Wide Web Conference*, Madrid, Spain, 2009, pp. 541-550.

[33] Balachander Krishnamurthy and Craig E. Wills, "Privacy diffusion on the web: A longitudinal perspective," in *Proceedings of the World Wide Web Conference*, Madrid, Spain, 2009, pp. 541-550.

Appendices

A. List of Sites in Each Category

Arts
youtube.com imdb.com deviantart.com movies.yahoo.com hulu.com scribd.com pandora.com last.fm veoh.com suite101.com

Employment
careerbuilder.com indeed.com monster.com jobsdb.com snagajob.com job.com quintcareers.com regionalhelpwanted.com vault.com employmentguide.com

Video Game News and Reviews
gamespot.com ign.com gamefaqs.com gametrailers.com gamesradar.com 1up.com gamespy.com gamershell.com eurogamer.net g4tv.com

Photo Sharing
flickr.com photobucket.com imageshack.us imagevenue.com webshots.com shutterfly.com pbase.com snapfish.com postimage.org kodakgallery.com

News
news.yahoo.com cnn.com nytimes.com weather.com huffingtonpost.com guardian.co.uk reuters.com timesofindia.indiatimes.com washingtonpost.com latimes.com

Travel
tripadvisor.com expedia.com travelocity.com orbitz.com southwest.com kayak.com travel.yahoo.com delta.com aa.com easyjet.com

Shopping
amazon.com ebay.com walmart.com ikea.com target.com bestbuy.com newegg.com overstock.com homedepot.com barnesandnoble.com

Relationships
match.com matchmate.ca singlesnet.com friendfinder.com blackpeoplemeet.com rsvp.com.au filipinaheart.com cupid.com afrointroductions.com connectingsingles.com

Generations and Age Groups
aarp.org thirdage.com kidzworld.com teenspot.com student.com golivewire.com kiwibox.com seventeen.com girlsense.com sugarscape.com

Sports
espn.go.com sports.yahoo.com cricinfo.com mlb.com goal.com foxsports.com cbssports.com nba.com nfl.com hattrick.org

Online Social Networks
twitter.com linkedin.com myspace.com livejournal.com orkut.com hi5.com bagdoo.com facebook.com stumbleupon.com tagged.com

Health
webmd.com mayoclinic.com mercola.com drugs.com menshealth.com medscape.com caloriecount.about.com weightwatchers.com kidshealth.org psychologytoday.com

B. Glossary

All the terms below are defined in greater detail in the paper. This section serves as a quick reference to the terminology used in the paper for the convenience of the reader.

Direct Leakage: Leakage during which one or more bits of private information is sent directly to a third party. This type of leakage includes leaking a user's name or address in some way.

Hidden Third Party: A third party with a sub-domain on a trusted site domain name.

Indirect Leakage: Leakage of data that is not in itself private information, but is information that can be used to access private information in some way. This type of leakage includes leaking a link to the user's profile.

Leakage of Private Information: A disclosure of private information to a third party without the explicit request of the user to do so.

Leakage During Signup: Leakage that occurs when a user is creating an account on a site.

Leakage of Search Terms: Leakage of terms used by a user when searching a site.

Leakage via Cookies: Leakage of information stored in cookies, generally due to a hidden third party.

Leakage via Referer: Leakage of information when information is put into the URL of a site and that site acts as a referer for an object stored on a third-party server.

Leakage via Request URL: Leakage of information that occurs when information is included in the URL of a request going to a third party.

Leakage via Title: Leakage of information that is being presented as part of a title of a page.

OSN: Online Social Network, such as Facebook and MySpace.

Private Information: Any information that either identifies an individual, can be linked to an individual, or that the individual would not like known.

Private Information Bit: A datum of private information that the user can reveal and may get leaked.