Privacy Diffusion on the Web: A Longitudinal Perspective

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

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

ABSTRACT

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

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

Categories and Subject Descriptors

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

General Terms

Measurement

Keywords

Privacy, Privacy Enhancing Technologies

1. INTRODUCTION

The European Union's Privacy directive [7] defines an "identifiable person" as "one who can be identified, directly or indirectly, by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity." It is well known that combinations of different data elements can lead to uniquely identifying a person. Privacy literature has introduced terms like de-identification (stripping identity information from data) and re-identification (ability to relate supposedly anonymous data with actual identities). Concerns about user privacy have risen dramatically with the increased dependence on the Internet for a wide variety of daily transactions that leave trails to be left in many locations. Web terms like cookies are widely known and modern browsers provide privacy protection choices.

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

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

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

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

Yet, there is little data about such privacy diffusion on the Internet resulting from individual user's actions involving visits to popular Web sites. In earlier work, we took a first cut at examining the problem of privacy diffusion on

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

the Web [13, 11]. In this paper, we present a longitudinal perspective of our study spanning four years exploring the nature and extent of tracking of user-related information by a large set of popular Web sites. Ours is the first such study to examine privacy diffusion over time that covers a broad set of technologies used for tracking and the potential for various measures against such tracking.

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

2. PRIVACY ELEMENTS

We now enumerate the list of privacy related data elements currently being tracked on the Web and the techniques used for such tracking. While our list is not exhaustive, we capture the most common elements and techniques.

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

Behavioral tracking is one of the oldest techniques employed on the Web. Behavioral tracking allows for monitoring user Web accesses across multiple unrelated Web sites. A common application is to see if a particular ad displayed on a site resulted in the user clicking on it. The common technique is to use a cookie that can be correlated across multiple sites; the aggregator knows that it is the same user who has visited these sites. The definition of a 'user' is somewhat nebulous: it could be simply the IP address present in the client HTTP request. But in combination with simple ambient information it may be possible to ensure that it represents a single user rather than multiple people sending requests from that IP address. For example, examining the access patterns over time, and the time periods and frequency of accesses, it may be easy to distinguish users even if multiple users are behind the same address. Web bugs (the 1x1 pixel GIF images) are another way to extract information about sites users are visiting. The advantage of behavioral tracking is thus the ability to create a profile of a user [16]. Use of tracking cookies is fairly ubiquitous [19] and there are known techniques to avoid them [22].

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

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

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

3. LONGITUDINAL STUDY

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

We also examine two important subsets of the broadly popular Web sites: a) consumer sites, where users do not just browse but supply additional personal information such as credit card numbers and b) fiduciary sites, where users

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

provide a variety of personal information including bank account numbers, and other personally identifiable information.

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

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

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

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

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

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

4. RESULTS

This section describes results from using the basic methodology for data gathering and analysis described in the previous section.

4.1 Longitudinal Results of Top Third-Party Root Domains

Focusing on the penetration of third-party root domains amongst the set of first-party servers in our basic test data set, Figure 1 first shows the cumulative penetration of the top-10 root domains at the time of each of the five epochs in our longitudinal study. The results show that the top-10 domains were used by 40% of first-party servers in Oct'05, but had extended to 70% of the first-party servers by Sep'08.

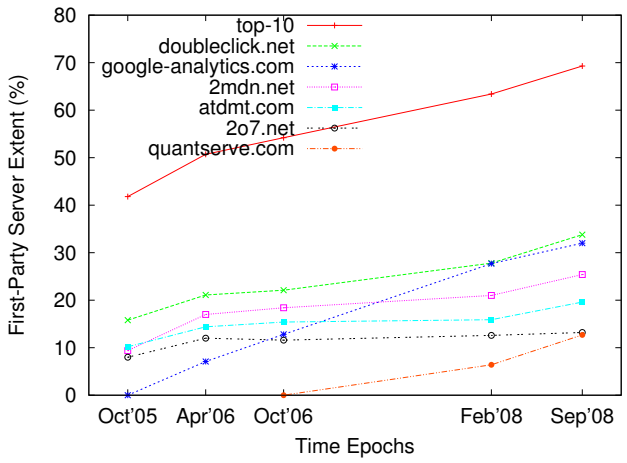

Figure 1: Extent of Top-10 Third-Party Root Domains in Each Epoch and Specific Contribution of Top Domains

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

4.2 What Are These Top Third-Party Domains Doing?

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

Rather than study all third-party domains, we focused on those with at least a one-percent penetration in a measurement epoch. Using this list as a starting point, we studied traces of requested objects, consulted the browser cookie database, and examined downloaded third-party JavaScript to better understand the nature of content served by servers in these domains. We primarily focused on the use of cookies by these third-party domains for tracking and whether these domains were using JavaScript to track users in conjunction with use of first- or third-party cookies. We found four types of third-party domains that track users amongst the set we examined.

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

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

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

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

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

4.3 Company Acquisitions

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

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

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

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

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

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

Figure 3 shows the growth of the Omniture family of domains. This family has grown through the increase use

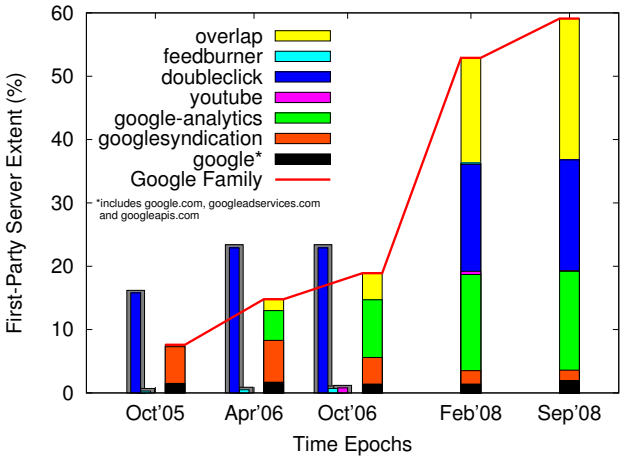

Figure 2: Growth of the Google Family

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

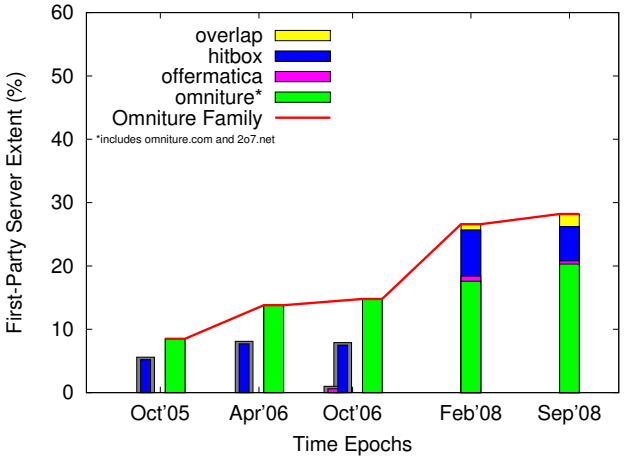

Figure 3: Growth of the Omniture Family

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

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

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

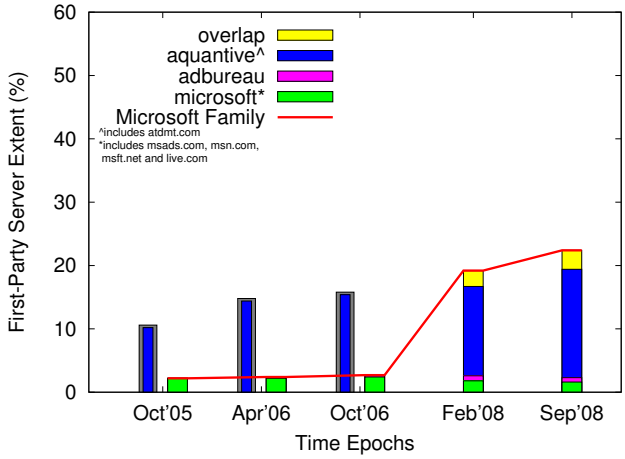

Figure 4: Growth of the Microsoft Family

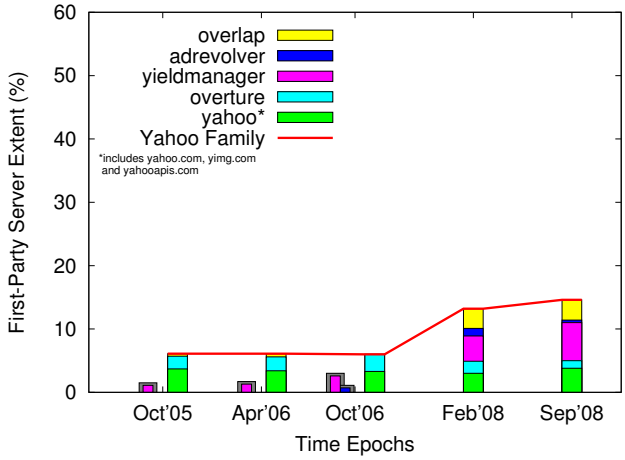

Figure 5: Growth of the Yahoo Family

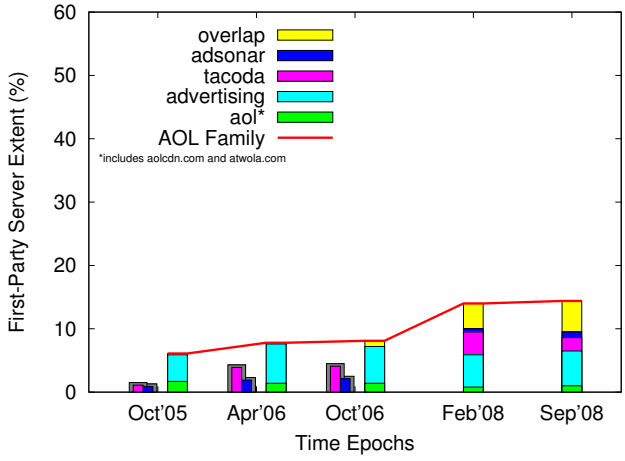

Figure 6: Growth of the AOL Family

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

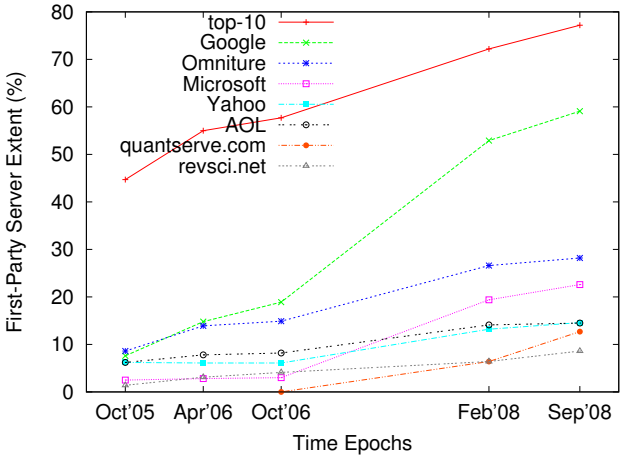

Figure 7: Extent of Top-10 Third-Party Families in Each Epoch and Specific Contribution of Top Families

4.4 Depth of Third-Party Penetration

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

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

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

The results show that the percentage of first-party servers with multiple top third-party domains has risen from 24% in Oct'05 to 52% in Sep'08. This increase has occurred despite the merger of previously independent domains through acquisitions. This increase is significant because it shows that now for a majority of these first-party servers, users are being tracked by two and more third-party entities.

4.5 Extent of Company Families in Consumer Sites

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

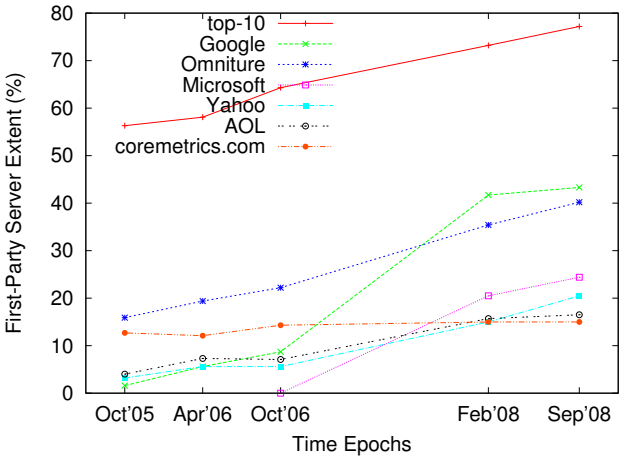

Figure 8: Extent of Top-10 Third-Party Families in Each Epoch and Specific Contribution of Top Families for Consumer Sites

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

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

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

4.6 Extent of Company Families in Fiduciary Sites

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

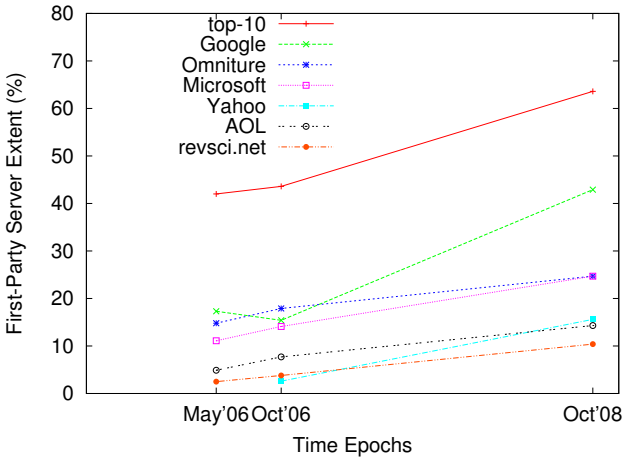

Figure 9: Extent of Top-10 Third-Party Families in Each Epoch and Specific Contribution of Top Families for Fiduciary Sites

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

4.7 Impact of Currency of Category Membership

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

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

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

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

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

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

5. LIMITATIONS OF PROTECTION TECHNIQUES

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

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

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

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

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

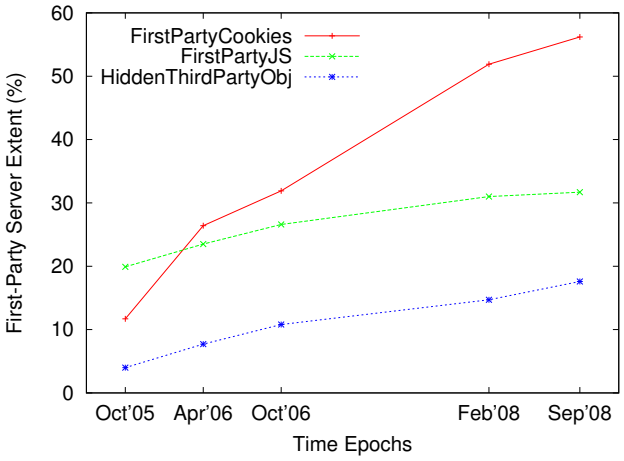

Figure 10: Growth of Hidden Third-Party Content

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

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

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

This game is also not limited to third-party sites doing analytics. Third-party sites doing behavioral tracking could deploy their content on what appear to be first-party domain servers and make use of first-party cookies to track users across first-party sites without any apparent use of third-party content or cookies. While we saw little evidence of this approach in our data, we would expect approaches like this to be used if enough users stop allowing third-party cookies.

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

6. DISCUSSION

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

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

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

The problem gets a bit more complex when we examine the popular free Web email services. These services require users to acknowledge that they accept a Terms of Services agreement, which spells out how a user's private information

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

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

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

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

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

7. SECONDARY PRIVACY DAMAGE

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

Earlier in Section 6, we referred to the potential construction of social graph by Web-based email services. Without the recipient's knowledge or consent, the communication between the first user (someone who has acceded to the Terms of Service) is available to the email service. If the recipient replies to the email then the contents of the response are also available without the second user ever being aware of the privacy policy of the email service.

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

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

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

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

8. CONCLUSIONS

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

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

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

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

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

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

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

9. ACKNOWLEDGMENTS

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

10. REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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