Generating a Privacy Footprint on the Internet

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

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

ABSTRACT

As a follow up to characterizing traffic deemed as unwanted by Web clients such as advertisements, we examine how information related to individual users is aggregated as a result of browsing seemingly unrelated Web sites. We examine the privacy diffusion on the Internet, hidden transactions, and the potential for a few sites to be able to construct a profile of individual users. We define and generate a *privacy footprint* allowing us to assess and compare the diffusion of privacy information across a wide variety of sites. We examine the effectiveness of existing and new techniques to reduce this diffusion. Our results show that the size of the privacy footprint is a legitimate cause for concern across the sets of sites that we study.

Categories and Subject Descriptors

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

General Terms

Measurement, Performance

Keywords

Privacy, Anonymity, Web

1. INTRODUCTION

Users on the Internet increasingly manage their daily interactions by accessing various Web applications that require them to supply private information such as credit card and bank account numbers. A necessary requirement on such sites is the safeguarding of all information that might be deemed as private to the users. Most users do not have an idea if any of the various bits of private information that add up to their identity is disseminated to parties other than the sites directly visited. The privacy implications of data gathered when users access Web sites needs to be examined closely.

Earlier [5] we examined non-primary content traffic (primarily advertisements) obtained as a result of visiting popular Web sites, and the resulting increase in objects, bytes and latency. Here, as part of constructing a *privacy footprint* metric measuring the dissemination of user-related information, we examine how browsing information related to individual users is tracked and aggregated across seemingly unrelated Web sites. We define the privacy footprint based on the set of sites visited by them, as the degree of interconnectedness seen through aggregator nodes. A large footprint indicates more privacy information leaked to aggregator nodes. Aggregator nodes in possession of information that can be tracked to individual users could potentially use it in a manner that violates the legitimate privacy expectations of users. Knowing the degree of potential leakage of private information may allow users to tailor their Internet activities, enable Web sites to be more circumspect about potential linkage of data, and allow for the emergence of new standards for protecting privacy.

Our privacy footprint metric can be computed in a straightforward manner, is augmentable over time, and comparable across individuals and organizations. Our goal goes beyond the issue of privacy: we can also measure the unwanted traffic involved in contacting third-party servers and the corresponding latency cost.

As a starting point, the study described in this paper, examines the rich set of interconnections between sites directly visited by the user and the additional sites caused to be downloaded as a result. Some of the downloads may be visible; many are not. The third-party sites visited indirectly often act as aggregators of information about the user's traversals through the Web. While some of the data gathered as a result is harmless, information about certain subsets of sites such as those related to managing personal fiduciary information (finance, health, insurance, mortgage etc.) raises stronger privacy concerns.

Our work is closely related to the concept of re-identification: the ability to relate supposedly anonymous data with actual identities. A collection of anonymous datasets can be combined with unanonymized datasets that were released separately in order to extract useful identification information. A canonical example is that of a dataset of medical records with just date of birth, gender, and geographical location information combined with another dataset of motor vehicle department which may have similar information. By merging the two, the more private information in the medical records leads to re-identification. In our study, the profile that could be captured by hidden nodes of a user's visit allows for such re-identification. If there is cookie information present then a hidden node, for example, could track a user who periodically visits a subset of their fiduciary sites with some predictable frequency. Such a profile could then be sold to other visible sites who may be interested in specific demographics.

The leakage of privacy is not a new concern and thus prevention techniques have been studied. Prevention of privacy leaks can be accomplished via blocking of unnecessary accesses to third party servers, use of intermediaries, etc. We thus study the role of how techniques used to block down-

loading of unwanted traffic can be combined with the goal of preserving privacy. The work described here is simply a starting point; the set of questions and concerns are broader and ongoing work will examine other aspects of the privacy issue.

2. STUDY

As a basis for investigating the privacy footprint for a set of Web sites we look at the connection between the directly (*visible*) visited servers and the *hidden* servers that are accessed as a by-product of visiting visible servers. Using a graph representation with nodes corresponding to servers, an edge connects a visible node to a hidden node when the access of the visible node causes the hidden node to be accessed.

As an example, when a user visits the page specified by `http://www.cnn.com/`, the server `www.cnn.com` is accessed along with servers `i.a.cnn.net`, `m.2mdn.net`, `m.doubleclick.net` and `cnn.122.2o7.net`. Different visible nodes often have an edge to the same hidden node, such as `m.doubleclick.net`, indicating that the server `m.doubleclick.net` is a potential aggregation point to track and correlate knowledge about a user's actions.

We say that visible nodes are *associated* with each other when they share one or more edges to a common hidden node. In some cases multiple hidden nodes within the same DNS domain are used. For example, the hidden nodes `cnn.122.2o7.net` and `dowjones.122.2o7.net` are part of the same `2o7.net` domain. We explore the impact of merging all hidden nodes with the same DNS domain. Another aspect is the characteristics of edges. Edges are assigned between visible and hidden nodes if at least one object is accessed, but the total number of objects is not important in terms of privacy. We distinguish between edges that lead to hidden nodes supplying cookies and those edges that do not.

Similar to [5], we start by gathering the list of all objects retrieved when a user visits a page specified by a URL. Extraneous content is often retrieved when Javascript is enabled. To gather realistic data about page downloads we used the Firefox browser augmented by the "Pagestats" Javascript extension [3], which records information about when each HTTP request was made and the response is received in an in-memory table and writing it out to a log file. The interface allows the extension to run the browser in batch mode where a list of sites is specified. The extension works well to efficiently and realistically retrieve over a thousand Web pages in a single batch. As in our previous work [5], we chose sites across various categories from Alexa's popular sites in the English language [2] with 100 pages in each of 13 different categories resulting in 1075 unique servers. These pages were retrieved from a single location in April/May 2006.

Since privacy has different connotations for different segments of users, we characterize information aggregation in tracking user activity across a broad range of Web sites. We also examine the specific role of cookies. We then narrow our examination to one important sub-category of sites: fiduciary sites involving personal financial information of users. We finally examine the effectiveness of methods to defeating tracking of users.

3. RESULTS

Our initial work on generating a privacy footprint for a set of pages focused on the dataset of popular sites from 13 Alexa categories. The pages in this set are served by 1075 servers (visible), which when accessed, cause an additional 2926 unique (hidden) servers to be accessed.

We first compute the number of associated visible nodes for each visible node to get an idea of connectedness in the graph. Two visible nodes are associated if they each are connected via an edge to a common hidden node. This "server" approach of using the server name for each hidden node fails to capture obvious organizational relationships amongst the hidden nodes. In our "domain" approach, hidden node servers with the same 2nd-level domain are merged into a single hidden node[1]. Visible nodes are not merged.

Using the 2nd-level name for combining servers within the same organization does not correctly capture all such relationships. Two frequently occurring hidden domain nodes are `google-analytics.com` and `googlesyndication.com`—nodes from the same organization, but not the same second level domain. We also found cases where what appeared to be a server in one organization (e.g. `lads.myspace.com`) was actually a DNS CNAME alias to a server (e.g. `lads.myspace.com.edgesuite.net`) in another organization (e.g. Akamai). We found these type relationships could be captured with an "adns" approach where all hidden nodes sharing the same set of authoritative DNS servers (ADNSs) were merged into a single hidden node.

To better understand whether this adns approach correctly groups servers of the same organization or if it leads to false positive errors, we examined the servers contained within the top-15 most frequently occurring ADNSs. The top-15 account for more than half of the ADNSs handling multiple servers. Doing spot checks on servers from these top-15 ADNSs using DNS lookup tools, WHOIS, traceroute and clustering analysis we observed an error rate of around 5% where servers from different organizations use the same ADNS.

Using the three approaches, Figure 1 shows a complementary CDF with the number of associations for all 1075 visible nodes. Along the y-axis, the results show that 61% of these visible nodes are associated with at least one other visible node using hidden nodes denoted by individual server names. When these hidden nodes are merged according to their domain then 72% of the visible nodes are associated with at least one other visible node, and when hidden nodes are merged according to their ADNS then this percentage grows to 82%. Along the x-axis, the results show a maximum of 338 (31%) associations for a single visible node under the server approach, a maximum of 443 (41%) associations under the domain approach and a maximum of 609 (57%) associations under the ADNS approach. The graph shows that over 60% of all visible nodes have associations with more than 100 other visible nodes using the ADNS approach. The breadth and the depth of these results indicates a significant number of relationships between popular Web sites visited by users that can be tracked via common, but typically hidden, servers in the Internet.

We next examine hidden nodes and the degree to which

---

[1]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.
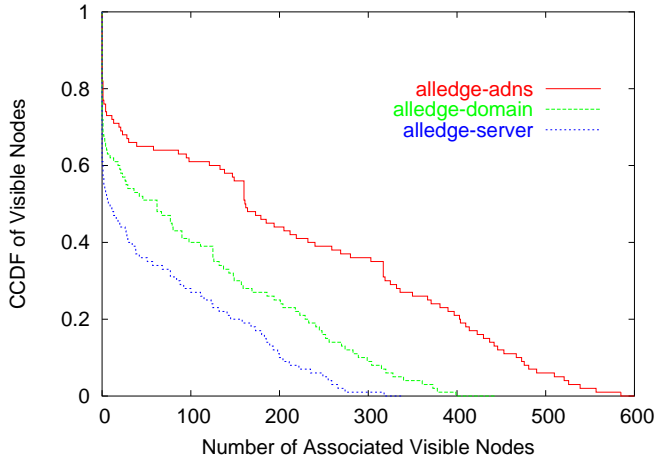
Figure 1: Complementary CDF for the Number of Other Visible Nodes Associated with Each Visible Node

visible node associations are concentrated amongst a set of hidden nodes. As an example of this concentration with the domain approach, Table 1 shows the top-10 hidden nodes in terms of the number of edges to visible nodes in our graph.

The top domain in the table, `doubleclick.net`, has edges to 19% of the visible nodes meaning that a user visiting a significant number of popular sites will likely download one or more objects from a server in the `doubleclick.net` domain. Other results in Table 1 show two separate domains with the name "google" and domain from Akamai. A CDN like Akamai is obviously in a position to correlate a range of sites visited by a user.

Table 1: Top-10 Connected Hidden Nodes Using Domain Approach

Hidden Node	Number of Connected Visible Nodes (%)
doubleclick.net	201 (19)
2mdn.net	185 (17)
atdmt.com	149 (14)
2o7.net	126 (12)
googlesyndication.com	91 (9)
akamai.net	80 (7)
google-analytics.com	78 (7)
hitbox.com	63 (6)
advertising.com	60 (6)
yimg.com	42 (4)

Table 1 shows the number of connections with visible nodes, but it does not accurately capture the cumulative effect of these connections because some visible nodes have associations with other visible nodes via more than one hidden node. For example, `www.cnn.com` and `online.wsj.com` are associated via both the `doubleclick.net` and `2o7.net` domains. Figure 2 shows the cumulative count of associations amongst visible nodes using a rank ordering of the hidden nodes for the three approaches for handling hidden nodes. All three results show a strong concentration of associations via the top hidden nodes. For example, the top-10

domain hidden nodes have edges to 559 (52%) of the visible nodes with associations. The top-10 ADNS nodes are connected to 682 (63%) of the visible nodes with associations. These results indicate that focusing on the top hidden nodes for analysis is appropriate.
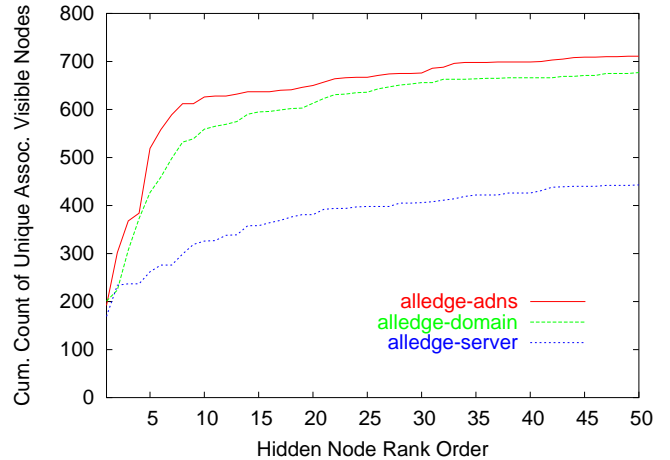

Figure 2: Cumulative Count of Unique Associated Visible Nodes Based on Hidden Node Rank Order

We were also interested in the specific URLs that are retrieved most frequently from the hidden nodes. These top-10 URLs are shown in Table 2 and consist of all JavaScript objects. While we did not specifically examine the code these objects contain, such examination would provide insight into how they work. Using techniques to block these URLs is also important to alleviate privacy concerns.

Table 2: Top URLs from Hidden Nodes

Hidden Node URL	Cnt
pagead2.googlesyndication.com/pagead/show_ads.js	89
www.google-analytics.com/urchin.js	76
rmd.atdmt.com/tl//DocumentDotWrite.js	63
m1.2mdn.net/879366/flashwrite_1_2.js	39
m.2mdn.net/879366/flashwrite_1_2.js	32
a.as-us.falkag.net/dat/dlv/aslmain.js	27
ar.atwola.com/file/adsWrapper.js	23
us.js2.yimg.com/us.js.yimg.com/lib/bc/bc_1.7.3.js	22
js.adsonar.com/js/adsonar.js	21
ar.atwola.com/file/adsEnd.js	21

3.1 Generating a Privacy Footprint

Thus far, we have shown that both the extent of associations between visible nodes and the concentration of these associations via hidden nodes is significant. We now summarize these measures on interconnectedness. We generate a "privacy footprint" intended to not only summarize the connectivity for a given set of sites, but to be used as a basis of comparison for different results. The footprint needs to capture the important metrics from the graphs in Figures 1 and 2. The metrics and their justification for inclusion in the footprint are:

1. The number and percentage of visible nodes with an

association to at least one other visible node, representing the breadth of associations amongst the set of sites.

2. The distribution (median, mean, max) of the number of visible node associations for a given visible nodes. This metric captures the CCDF of Figure 1 and represents the depth of the associations—do visible nodes have relatively few or many associations?

3. The contribution of the top-n rank ordered hidden nodes to these associations. Based on Figure 2, we examine the top-10 hidden nodes to understand the degree of association concentration because of these nodes.

To illustrate this privacy footprint and how it can be compared with other privacy footprints, we generated the footprint with these metrics for the data gathered in April 2006 for the results presented above and for data gathered in October 2005 for previous work [5]. Data gathered each time is for the same set of sites. The respective privacy footprints for each timeframe and each approach for merging hidden nodes are shown in Table 3.

Table 3: Privacy Footprint of 1075 Alexa Web Sites (April 2006 and October 2005)

Timeframe/ Approach	Visible Nodes w/ Assoc's (%)	Number of Assoc's for a Visible Node			Assoc's Via Top-10 Hidden Nodes
		Med.	Mean	Max	
apr06/adns	879 (82)	225	247	609	682
apr06/domain	779 (72)	125	144	443	559
apr06/server	659 (61)	82	103	338	378
oct05/adns	853 (78)	121	170	527	585
oct05/domain	718 (66)	80	98	347	456
oct05/server	591 (54)	31	69	261	333

The first three lines in Table 3 simply summarize data in Figures 1 and 2 with the last three lines in the table showing the same metrics for data gathered six months earlier for the same set of sites. The results show that while the number of visible nodes with associations has increased roughly 5% across the three approaches, the metrics for the number of associations for a visible node have increased roughly 50% for the mean indicating a significant increase in the associations via the hidden nodes. The concentration of associations among the top-10 hidden nodes has increased roughly 20% for the adns and domain merger approaches. The large increase in these metrics indicates a growing potential to track and correlate user activity across seemingly unrelated Web sites on the Internet.

3.2 Global Study

We also examined the use of our privacy footprint for a much larger and more diverse set of Web sites using the "Top-100" sites identified by Alexa for 68 countries and 19 languages around the world [2]. The results indicate that diffusion of potential privacy information is an issue for users of sites around the world. Table 4 shows hidden nodes that appear in at least 15% of the per-country top-10 hidden node lists for the 68 countries. The most frequently occurring hidden node in the per-country top-10 lists is `google-analytics.com`, which appears in 90% of

the lists. The nodes in Table 4 are similar to those in Table 1 with additions such as `yahoo.com`, `statcounter.com` and `imrworldwide.com`.

Table 4: Hidden Nodes Appearing in 15% of 68 Per-Country Top-10 Lists

Hidden Node	Number of Appearances in Country Top-10 Hidden Node List (%)
google-analytics.com	61 (90)
yahoo.com	58 (85)
yimg.com	47 (69)
googlesyndication.com	44 (65)
doubleclick.net	39 (57)
2o7.net	31 (46)
atdmt.com	24 (35)
2mdn.net	22 (32)
statcounter.com	15 (22)
imrworldwide.com	14 (21)
adbrite.com	14 (21)
webstats4u.com	10 (15)
ratteb.com	10 (15)

3.3 Impact of Cookies

Cookies are a common mechanism for Web sites to maintain state during e-commerce transactions or maintain personalization context for a user. Cookies are also used by tracking servers to more accurately identify a user as the user navigates between different Web sites. If pages from these Web sites cause objects to be retrieved from the same tracking server and this server has a cookie associated with it then the server receives this cookie on each retrieval.

Hidden nodes in our study that have cookies associated with them are particularly troublesome for privacy. To analyze the impact of cookies, we gathered whether or not cookies are associated with a server during data collection. This data gathering was done by configuring the browser to accept all cookies and then to harvest the `cookies.txt` file (maintained by Firefox as part of a user's profile) after a set of pages had been retrieved. The text file contains one line for each cookie with the server (or domain) as the first field. For our analysis, we do not care about cookies set by servers of the visible nodes nor do we care how many cookies are set—one cookie suffices for privacy leakage.

We used the cookie data to modify our graphs to include only edges that are connected to hidden nodes that have cookies associated with them. We then recomputed the privacy footprint metrics for each of the merger approaches on the Alexa dataset with results shown in Table 5. For convenience, Table 5 repeats the all edges results reported in the April 2006 results of Table 3. We also drop inclusion of the server approach as it does not merge all servers of an organization.

The metrics for edges with cookies in Table 5 are smaller than comparable metrics where all edges are used, but these associations all have cookies attached to the object requests. The top-10 hidden nodes with cookies attached for the domain approach are shown in Table 6. These nodes are collectively responsible for connections to 483 distinct visible nodes as shown in the last column of Table 5. Note the difference between 483 and the summation of counts in Table 5 is because visible nodes are associated with other visible nodes via multiple hidden nodes.

Table 5: Privacy Footprint of 1075 Alexa Web Sites for All Edges and Those with Cookies

Edges/ Approach	Visible Nodes w/ Assoc's (%)	Number of Assoc's for a Visible Node			Assoc's Via Top-10 Hidden Nodes
		Med.	Mean	Max	
alledges/adns	879 (82)	225	247	609	682
alledges/domain	779 (72)	125	144	443	559
cookie/adns	604 (56)	186	205	578	503
cookie/domain	595 (55)	148	145	392	483

Table 6: Top-10 Connected Hidden Nodes with Cookies Using Domain Approach

Hidden Node (Domain)	Number of Connected Visible Nodes (%)
doubleclick.net	201 (19)
atdmt.com	149 (14)
2o7.net	126 (12)
hitbox.com	63 (6)
advertising.com	60 (6)
tacoda.net	40 (4)
revsci.net	32 (3)
webtrendslive.com	28 (3)
falkag.net	27 (3)
yahoo.com	26 (2)

3.4 Fiduciary Sites

Having applied our methodology to a broad set of sites, we next examined potential sharing of information about access to sites that manage personal fiduciary information. Users provide private information such as credit cards and bank account numbers to such sites. We constructed nine categories of such sites: credit, financial, insurance, medical, mortgage, shopping, subscription, travel and utility. We identified 81 sites across these nine categories with the specific sites for each category.

From a privacy standpoint, it is vital to reduce diffusion of information about access to these categories of sites. We did not actually login to any of these sites for our testing, but we assume that users would be most likely to visit the home page of each site before logging into a site (or being identified based on cookies). An interesting piece of future work would be to examine the diffusion of access information after login to a site has occurred.

Results for the privacy footprint across these 81 sites are shown in Table 7. The size of the privacy footprint is generally smaller than the Alexa dataset, both in terms of the number of associated visible nodes and the distribution of associations of these visible nodes. Although not shown, the top hidden nodes are similar to what we found for the Alexa dataset with domains doubleclick.net, atdmt.com and 2o7.net as the most connected hidden nodes.

We also looked at the privacy results for these sites using the categories for each site. Our concern is that a user could have fiduciary interests with a site in each category and would be particularly concerned if accesses to different categories of sites could be tracked. We found that in terms of privacy none of the 10 visible nodes in the medical category had any associations with other visible nodes in our dataset. This is a good result in terms of privacy concerns.

Table 7: Privacy Footprint of 81 Fiduciary-Related Sites

Edges/ Approach	Visible Nodes w/ Assoc's (%)	Number of Assoc's for a Visible Node			Assoc's Via Top-10 Hidden Nodes
		Med.	Mean	Max	
alledges/adns	52 (64)	11	11	32	40
alledges/domain	41 (51)	6	7	25	32
cookie/adns	47 (58)	10	10	32	38
cookie/domain	37 (46)	7	7	20	30

It is also possible to construct a set of nine sites, one from each category, where no site has an association with another site. A hypothetical user, whom accesses this particular set of sites would have no privacy concerns. However, at the other extreme we found it is possible to construct a set of sites, one from each category, where a site from each of the non-medical categories is associated with a distinct site in at least one other category with a mean of five and a maximum of six associations with sites in other categories. This result is consistent whether or not the presence of cookies is considered. It indicates that it is possible for information across these categories to be shared.

3.5 Methods to Defeat Tracking

Given the widespread use of hidden nodes that have the potential to track the browsing behavior of users across a large number of visible nodes, the last question we investigate is what retrieval methods can be used to defeat such tracking. In this section we discuss two such methods and examine their effectiveness in terms of reducing the privacy footprint. We use the Alexa dataset used for initial work presented previously in this section.

3.5.1 Ad Blocking

We can block objects used for tracking by treating them as extraneous content (such as advertisements). Using the same methodology as in [5], we use the Adblock Firefox extension [1], which blocks the retrieval of objects whose URL match one or more pattern rules specified by the user. Likewise, we use a ruleset named "Filterset.G" [4], which is commonly accepted as best practice for using Adblock to block extraneous content. For this analysis we used the 2006-03-08 Filterset.G ruleset version and converted the rules to Perl regular expressions (Perl and JavaScript use the same regular expression syntax) and filtered out all objects matching at least one rule.

Table 8 shows the privacy footprint results for the Alexa dataset with all URls that match a Filterset.G rule being filtered out. Comparing the alledges results in this table with those in Table 5, we see that the mean number of associations for a visible node using the adns merger have dropped roughly 50% and have dropped roughly two-thirds for the domain merger approach. The cookie-only based results in the lower-half of Table 8 show a more significant drop when compared to cookie-only results in Table 5.

These results indicate that ad blocking techniques can significantly reduce the potential tracking by hidden nodes, but not eliminate it. Filtering eliminates many of the domains shown in Tables 1 and 6, but objects from domains such as 2mdn.net, revsci.net and webtrendsalive.com are not filtered.

Table 8: Privacy Footprint of 1075 Alexa Web Sites Using Adblock with Filterset.G Rules

Edges/ Approach	Visible Nodes w/ Assoc's (%)	Number of Assoc's for a Visible Node			Assoc's Via Top-10 Hidden Nodes
		Med.	Mean	Max	
alledges/adns	795 (75)	91	119	399	506
alledges/domain	595 (56)	27	50	227	327
cookie/adns	288 (27)	38	83	343	196
cookie/domain	274 (26)	17	17	55	183

The adns footprint in Table 8 is larger across all metrics than the domain footprint because the adns approach combines servers that are syntactically distinct, but share the same set of ADNSs. The largest ADNS node when considering cookie-only results is one connected to yahoo.com sites as well as sites in the burstnet.com domain. The second largest ADNS set is for an Akamai ADNS serving objects with cookies for different visible nodes.

The results show that standard ad blocking techniques improve on, but do not eliminate, privacy concerns. The reasons for these mixed results are that not all objects used for tracking may appear to be an "ad" and that not all associations among servers used for tracking may be evident based solely on a server name.

3.5.2 Blacklisting Top Hidden Nodes

The previous results show that blocking ads may not be the best approach for reducing tracking. A more direct approach is to identify the most frequently-used hidden nodes and simply block retrieval of all objects from these nodes. Table 9 shows the results of applying this direct approach where all objects belonging to one of the top-10 hidden node domains shown in Tables 1 and 6 are filtered.

Table 9: Privacy Footprint of 1075 Alexa Web Sites Using Blacklist of Top-10 Hidden Nodes

Edges/ Approach	Visible Nodes w/ Assoc's (%)	Number of Assoc's for a Visible Node			Assoc's Via Top-10 Hidden Nodes
		Med.	Mean	Max	
alledges/adns	811 (75)	44	108	415	450
alledges/domain	604 (56)	14	17	90	205
cookie/adns	374 (34)	32	84	392	234
cookie/domain	359 (33)	18	20	73	195

The results in Table 9 show small variations with the Adblock results in Table 8, but overall the footprint results are similar. The results do not improve because even with the top hidden nodes filtered out with the blacklist method, there are still many other hidden nodes to interconnect the visible nodes. This result indicates that visible nodes are often associated via multiple hidden nodes.

4. CONCLUSIONS

Privacy is a central concern of users in the Internet and this work examines one privacy issue—the potential to track and correlate knowledge about a user's actions across seemingly unrelated Web sites. We used an approach that defines edges between the "visible" nodes, which are the servers that users directly access, and the "hidden" nodes, which are the servers that are accessed as a result accessing a visible node. We use this approach as a basis to define and construct a *privacy footprint*, which monitors the diffusion of information about a user's actions by measuring the number of associations between visible nodes via one or more common hidden nodes.

The privacy footprint metric can be computed in a straightforward manner, augmentable over time, and is comparable across individuals and organizations. The openness of the Web, the flexibility and extensibility of modern browsers like Firefox, allow us to construct tools that can carry out measurements concurrent with normal browsing by the user.

Using the footprint for a set of popular sites, we found that the mean number of associated sites has increased by 50% in the past six months. This is a significant increase in a relatively short time. Narrowing our examination just to sites that supply cookies indicates that the privacy footprint is still extensive. Our results show that the size of the privacy footprint is a legitimate cause for concern across all sets of sites that we studied.

We found that methods such as ad blocking and blacklisting of hidden nodes to defeat tracking of user actions across Web sites are only partially effective due to difficulties in identifying all hidden nodes and in identifying organizational dependencies amongst these nodes.

For future work, we believe an alternate approach to consider is a "filter-in" technique, which by default whitelists servers in the domain of the visible node. This technique is simpler to specify and, based on preliminary investigation, more effective compared to other methods in limiting the privacy footprint. However, we need to examine usability concerns of this technique and we found it needs to be augmented with a whitelist of allowed hidden nodes as well as knowledge of hidden nodes associated via shared ADNSs. We plan to build or extend an existing browser extension to provide this functionality.

In conjunction with this technique, we plan to pursue development of an extension to actively monitor and alert the user of any associations made between visible nodes as a user browses the Web. In this work, we gather information and then perform off-line analysis to determine associations. An extension that performs this work in real-time would both be valuable for users to understand the spread of information about them and could be used as input for filtering rules.

Finally, the definition of a privacy footprint provides us a basis on which to continue to monitor the diffusion of privacy information. We plan to do so for popular sites as users and content providers adapt their approaches in this important domain.

5. REFERENCES

[1] Adblock. http://adblock.mozdev.org/.
[2] Alexa: Most popular web sites. http://www.alexa.com/.
[3] Scot DeDeo. Pagestats, May 2006. http://www.cs.wpi.edu/~cew/pagestats/.
[4] Official home of Filterset.G. http://www.pierceive.com/.
[5] Balachander Krishnamurthy and Craig Wills. Cat and mouse: Content delivery tradeoffs in web access. In *Proceedings of the International World Wide Web Conference*, Edinburgh, Scotland, May 2006.