

A User's Perspective on Privacy and the Web*

Anna Belle Leiserson**

Troubled by the mystery of "cookies" and how they affect privacy on the Web, Ms. Leiserson digs deep to explore not only cookies but also other technologies that can be used to breach a Web user's privacy. She suggests practices a user can follow to minimize their impact.

¶1 Over the past few months I have been keeping an eye on privacy and the Web. In part this is because I have moved into an information technology position in my law school and thus am inundated with notices about security breaches. But it's also because of long-felt qualms about "cookies" that I share with many librarians. Such qualms are typically fueled by a lack of understanding of the technology, and thus my mission in this article is to identify the relevant technologies, dissect them, and then translate my findings from technobabble into English.

¶2 An important part of defining these technologies is first clarifying what they are not. This article will not be an overview of privacy and computing in general, much less an analysis of Internet security. Nor will it be a political or legal analysis of Web privacy, although understanding these technologies is a helpful foundation for such analyses. Just to clarify, here are some issues that will not be covered:

- **Viruses.** While it is possible to catch a virus such as the Nimda worm from a Web site, these are much less common than e-mail-borne viruses and fall under the larger umbrella of security.
- **Spyware.** This comes in two basic varieties: hardware and software agents. The hardware includes amazing devices, such as KeyGhost, that can be secretly attached to computers and capture all keystrokes. This is obviously beyond the Web. The software is a bit closer to our topic, and can come in particularly insidious forms that are bundled with other software. A good example is LimeWire, the popular file-sharing freeware that, unbeknownst to its users, was (until recently) also tracking their Web surfing patterns. However, this too is arguably a general computing technology.
- **Denial of service (DoS) attacks.** These have shut down major Web sites, including Yahoo! and Amazon.com, but are security, not privacy, issues.

* © Anna Belle Leiserson, 2002.

** Webmaster, Vanderbilt University Law School, Nashville, Tennessee.

- **Software “holes.”** Software flaws that hackers can take advantage of abound. Web browsers, particularly Internet Explorer, are among the more common types of software that need patching. However, I would term this more of a software, and thus computing, issue.

¶3 Such issues are the making of whole careers and enterprise-level efforts, even at behemoth Microsoft. So instead we will remain focused simply on the Web and privacy, particularly with an eye to identifying good user practices.

¶4 As I have looked more methodically at these unpleasant technologies and then sifted through them, I have discovered to my surprise that the lowly cookie remains the focal point of Web privacy issues. However, there is a bit more to it than simply cookies, so let us back up and begin by breaking down the categories of “tracks” that users leave on the Web.

Log Files

¶5 At the most basic level, whenever users surf the Web, they automatically leave a simple list of all pages requested in “log files” (or “access logs”). While I discussed log files in an earlier article,¹ that was from a different perspective. To review, a log file is data stored on all Web servers. In a log file, every access is one record, and every record typically has the following fields:

- Date and time
- File accessed, including all Web pages (e.g., html files), graphics, PDF files, and anything else a user pulls up with a request to browse a particular portion of a particular site
- Referring page, i.e., from whence the user came to this particular file
- The browser type and version
- The platform
- Most important to privacy, the IP number of the machine. This is the number assigned to the computer so that it can access the Internet. An example would be 198.252.9.188. The first part of this number identifies the network, in this case Washburn University. The rest is assigned locally and identifies the particular machine. The scheme behind this numbering is similar to ISBNs. However, while it is possible to identify machines with this information, it is more difficult than it is for ISBNs. Thus log analysis tools only analyze the network, not the machines. But for those wanting to understand privacy and the Web, this is an important concept. The machine leaves an individual record, and it is possible (though cumbersome and difficult, and thus unlikely) for a Web server administrator to determine who was looking at what part of their site when.

1. Anna Belle Leiserson, *Web Wizards: Engineers, Artists, and Librarians*, 94 LAW LIBR. J. 167, 170, 2002 LAW LIBR. J. 11, ¶ 13.

¶6 To show how a log file constitutes a “track” left by a user, let’s consider the log file of my law school’s Web site, which on a slow day recently had approximately 20,000 lines (or records). One of those lines reads:

```
129.59.1.10- - [31/Mar/2002:10:02:33 -0600] "GET /library/research/legal_writing.html
HTTP/1.0" 200 7359 "http://www.google.com/search?hl=en&q=oral+argument+conclu-
sion+law+school++sample" "Mozilla/4.0 (compatible; MSIE 6.0; Windows 98; Win 9x
4.90)"
```

To translate, this means that on March 31, 2002, at 10:02 A.M., someone went to our library’s “Research Guide on Legal Writing” (/library/research/legal_writing.html) using a search query from Google (<http://www.google.com/search?hl=en&q=oral+argument+conclusion+law+school++sample>). The person was using Internet Explorer 6 on a Windows 98 machine. The number assigned to that machine (which I have actually changed for the sake of privacy) was 129.59.1.10. The 129.59 will tell the log analyzer that this user’s network was Vanderbilt University.

¶7 In trying to understand logs and Web privacy, the most important concept is that while it is often possible to determine individual machine use, the log files are an enormous data quagmire. Thus Web administrators almost never look for individual pieces of information. Instead they use number-crunching software that only analyzes at the network level. It is possible, however, for things to happen that might affect an individual’s privacy, such as log records being subpoenaed. This trail of data is analogous to a library’s circulation records, and thus issues of log file privacy and protection of borrowing records are similar.

Form-Based Information and Encryption

¶8 While a user could easily be unaware of the tracks left in a log file, a much more obvious track is left whenever one fills out an interactive form on the Web. An interactive form is typically a Web page with a series of boxes where a user types in the requested information and then submits it. What happens when one submits form data? There are a range of possibilities, but typically one of four things occurs: the data are e-mailed to a designated e-mail box; it is stored in a file; it is stored in a database; or some combination of these options occurs. Data stored in databases can and do become massive repositories of consumer information.

¶9 When the forms request sensitive data such as credit card or social security numbers, privacy become an even larger issue. The data ought to be on a “secure site,” but in fact a Web form requesting such information can easily be put on a regular (unsecured) site—presumably by a webmaster who does not know what he or she is doing. Secure sites typically use SSL (“Secure Sockets Layer”), a protocol that encrypts and decrypts data, making it much harder for hackers to read the information as it travels between computers. Whenever such information is requested, there are various ways a user can be sure the site is secure. One way is

to simply look at the URL. If it starts with “https” instead of just “http,” it is secure. Also, most of the current major browsers (Internet Explorer, Netscape, and Opera) have a small padlock near either the lower right or upper left corner of the browser. It will either appear or will close when on a secure page.

Cookies

¶10 A cookie is a small amount of data a Web site’s server stores on the user’s (client’s) hard drive. Its purpose is to enable the server to remember particular pieces of information about the user. In other words, cookies are the main way to personalize Web site viewing. However, Web site personalization and privacy are often at loggerheads.

¶11 It is somewhat ironic that this technology is named a “cookie.” In my world of the chocolate chip and oatmeal raisin variety, it is easy to find out the ingredients in a cookie. That is far from the case with Web cookies. So to investigate cookies, I started by creating my own.

¶12 There are a variety of ways for a webmaster to create cookies, none of which can be seen by a typical user, even a user who likes to use the browser “View Source” command. Cookies are coded at the start of a Web page, using a scripting language such as JavaScript, Perl, ASP, or my choice, PHP. My script was quite benign, not to mention simple. It merely changed the background color of the Web page. The critical line read:

```
setcookie (“BGColor”, “”, time()- “100”, “”, “”, “”);
```

Note that there are six fields, separated by commas in the parentheses, after the “setcookie” command. These are the six basic parameters of a cookie: the name, value, expiration date, path, domain, and whether or not it is a secure connection. Mine was so simple it only had a name and an expiration time of one hundred days. This means whatever color I set for this page would appear again every time I went to this site using the same browser for the next one hundred days—unless of course I changed the settings, in which case it started anew with a different color.

¶13 Cookies are browser specific, however. Thus, if I switched browsers, the new one would know nothing about the previous settings. The browser used also determines where the cookies are stored and the level of user control over the cookies.

¶14 Internet Explorer 5.5 and 6.0 on Windows 2000 store cookies individually. They are found in the “Documents and Settings” folder. For every user of a given machine, “Documents and Settings” will have a folder with the username, and that folder in turn will have a folder named “Cookies.” So my cookies can be found in: C:\Documents and Settings\leiserson\Cookies. Once in this folder, the cookie can be found by looking for the current date. This cookie was named: “leiserson@

learning[2].txt.” I could open it with any text editor (such as NotePad or Microsoft Word). In the browser, I had set the background color to be green, so here is what the cookie file contained:

```
BGColor
green
localhost/learning/
1024
2245291264
29717016
3782706240
29484185
*
```

¶15 Netscape 6 for Windows also stores cookies in the Documents folder, but it buries them even deeper. Thus mine was found in: C:\Documents and Settings\leiserson\Application Data\Mozilla\Profiles\default\r69krrqy.slt\ Netscape continues to use the original method of storing cookies, with just one file named “cookies.txt”. Here is the same cookie data in Netscape, stored in cookies.txt as one line separated by tabs:

```
localhost    FALSE    /learning    FALSE    1118889682    BGColor    green
```

Part of why Netscape has buried this file is because it is a really bad idea to edit this file by hand.

¶16 However, with version 6, Netscape has added excellent cookie management tools. Going into the “Cookie Manager,” I could also see that site “localhost” had a cookie name of “BGColor” with information “green” in path “/learning,” a secure server value of “no,” and an expiration time of June 1, 2005. In other words it translates all of the above mumbo jumbo. I could also remove this cookie with a click of the button or remove all cookies.

¶17 When my browser (be it Netscape or Internet Explorer) contacted this server (“localhost”), the data in the relevant cookie was all it saw. It is not much. My next step was to compare this cookie data to that used by a highly personalized site. I chose Amazon. Amazon’s cookie looked quite similar, the big difference being the information (or value) field. Instead of “green,” the value was a string of thirty-two random characters. It did not include my e-mail address or any other sensitive data. Of course that string has meaning to Amazon’s computers. Presumably it keys into their databases, where they do have a record of my e-mail address and many other things, such as all of the pages at their site that I have looked at recently.

¶18 This is a quick demonstration of the power of cookies and the wide variety of uses to which they may be put. At one end of the spectrum, they can be used for very simple forms of personalization, devoid of any significant information about the user. At the other end, they can be used to instantly tap into and expand a huge repository of information the user has already provided.

¶19 A boon to users is that newer browsers, such as Internet Explorer 6 and Netscape 6, can be customized to accept or reject cookies from particular sites on a routine basis. This is done by setting the browser's preferences to prompt (or warn) before accepting cookies. From these same settings, it is also possible to routinely reject all "third-party cookies." Third-party cookies are typically the products of advertising agencies. They work the same as other cookies. The difference is that users encounter one site's cookie at another site. Doubleclick.net is a particularly pervasive example of third-party cookies. Users with any cookie files at all on their hard drives are likely to have Doubleclick among them. Most have Doubleclick cookies without ever having visited Doubleclick's site. Doubleclick is a targeted marketing company, and its cookies are encoded in thousands of Web sites. Imagine surfing habits captured from countless Web sites and stored in massive databases, and it is easy to see how such information can be very effectively used and abused.

Posting to the Web

¶20 Another privacy issue is the vague category of information about a user that is directly uploaded to a Web site. The most common example is e-mail addresses on the Web. Not only do companies capture these e-mail addresses to use for spam, there is "e-mail harvesting" software designed to automatically scan thousands of Web pages for e-mail addresses and extract them for easy bulk mailing.

¶21 There is also the issue of archived e-mail posted to the Web. There are various search engines that will find these postings. This not only leaves surprise content scattered around the Internet, it is also another way for spammers to gather e-mail addresses.

¶22 There are also more random postings such as résumés uploaded to job sites like Monster.com. According to an article on the Privacy Foundation's Web site, "[r]ésumés may be stored by online job sites for many years, and may be misused for data mining and even identity theft."² Certainly it is possible to stumble across information one thought was deleted from the Web long ago by using the Wayback Machine of the Internet Archives.³

Counter Measures

¶23 While studying this underbelly of the Web creates a bleak picture, the news is by no means all bad. According to a recent report of the Progress and Freedom

2. Pam Dixon, *Click, You're Hired. Or Tracked . . . A Report on the Privacy Practices of Monster.com*, at www.privacyfoundation.org/privacywatch/monster.asp (Sept. 5, 2001).

3. INTERNET ARCHIVES, WAYBACK MACHINE, at www.archive.org/ (last visited Apr. 30, 2002) ("The Wayback Machine makes it possible to surf pages stored in the Internet Archive's web archive.")

Foundation,⁴ Web sites are collecting less personal data from users than in the past. A survey of one hundred of the most popular sites, such as Amazon.com and Google, indicated that the sites gathering such information had dropped from 96% to 84%, and the sites using third-party cookies from 78% to 48%.

¶24 In addition, the report noted that the number of sites with privacy policies is increasing. Privacy policies actually come in two very different forms. The most common type is a Web page somewhere on the site that describes its policies on the use of information collected from visitors to the Web site. These statements might be anywhere on the site and both how they are linked to and what exactly they cover is left to the judgment of the site itself.

¶25 The second type of privacy policy is consistent and standards-based, but it is also very new, quite rare, and considerably more difficult for webmasters to generate. It uses the Platform for Privacy Preference Project (P3P) of the World Wide Web Consortium. P3P provides users “a standardized set of multiple-choice questions, covering all the major aspects of a Web site’s privacy policies. Taken together, they present a clear snapshot of how a site handles personal information about its users.”⁵ Currently Internet Explorer 6 (IE6) is the only browser to have implemented P3P. In IE6, a user may click on “View,” then “Privacy Report,” and “Summary,” and if the Web site has a P3P policy, an abstract appears, providing the site’s answers to such questions as why the information is collected, who has access to the information, and how long the information is retained. The browser will also compare this with the user’s stated privacy preferences.

Good Practices and the Savvy Web User

¶26 “Privacy is dead,” according to an oft-quoted remark by Scott McNealy, CEO of Sun Microsystems. Out of context, this an overstatement, but for Web users, privacy has always been an illusion. What is changing is our understanding of just how much our privacy has or has not been invaded, and how to manage this reality.

¶27 The best way to avoid leaving tracks is simple—don’t use the Web. Of course, this is a particularly unacceptable solution for librarians. Slightly less drastic methods would be to use only public kiosks to access the Web or to rely on anonymizers.⁶ These too are relatively unpalatable ways of protecting privacy for librarians.

4. News Release, Progress & Freedom Foundation, On-line Survey Shows Progress on Privacy (Mar. 27, 2002), *available at* www.pff.org/pr/pr032702privacyonline.htm.

5. WORLD WIDE WEB CONSORTIUM, PLATFORM FOR PRIVACY PREFERENCES (P3P) PROJECT, *at* www.w3.org/P3P/ (last revised Apr. 22, 2002).

6. A word about anonymizers. These are defined as services that make it possible for users to visit Web sites without allowing the sites to gather information about them. While cookies are now the devil one knows, anonymizers vary and are probably the devil one does not know. I am particularly wary of them, having picked up a virus from one of the best known of these services.

¶28 However, even without going to these extremes, savvy Web users can achieve adequate privacy. By “adequate privacy” I mean Web use that minimizes the tracks left, protects sensitive data such as social security and credit card numbers and even e-mail addresses, and most of all, allows the user to be knowledgeable about who might be watching and how.

¶29 Thus we come to the ultimate question: what are reasonable precautions to use when surfing the Web? To answer, let me conclude with a list of good practices for improving Web privacy.

- Use a recent browser, such as Netscape 6 or Internet Explorer 6, that has a good cookie manager
- Use high-privacy settings in the browser
- Read privacy statements before entering sensitive data
- If entering sensitive data, be sure the site is secure
- Monitor where your e-mail address is posted on the Web
- Remember when posting to e-mail lists or directly to the Web that this information could resurface later in unexpected places on the Web
- Most of all, be cautious in what you tell a Web site about yourself