SPECIAL FOCUS: DATA AND INFORMATION COLLECTION ON THE NET

The Anonymizer

Protecting User Privacy on the Web

by Justin Boyan

Introduction

In a well-known New Yorker cartoon from July 1993, a dog sitting comfortably at his computer terminal says smugly, "On the Internet, nobody knows you're a dog." Surfing the web certainly does feel anonymous, like listening to the radio or browsing in a bookstore. In fact, web surfers leave identifiable tracks at every web site they visit.

Privacy and anonymity are important to many people for many reasons:

- The employee or politician who wants to protect his or her privacy when viewing sensitive medical information, a competitor's web site, sexual materials, or a web site catering to a marginalized group (e.g., gay rights, pro-choice or pro-life).
- The scientist who is asked to anonymously review a colleague's article submission and wants to gather background materials from the author's web site.
- The law-enforcement agent who wants to investigate a web site suspected of criminal activity without revealing that his or her Internet host is, say, `fbi.gov`.
- The consumer who wants to prevent marketers from compiling user profiles of his or her browsing, newsreading, shopping, financial and travel interests.

As a matter of principle, "the American tradition has maintained not only the right to print and speak freely, but also the right to read anonymously" (Camp, 1997). Since the web serves as not only a library but also a virtual meeting place for groups of all kinds, the right of free association without surveillance also applies to users of the new medium.

In this article, I first discuss the technology which allows web surfers' privacy to be violated. I then outline the design of my "Anonymizer" system which protects against such privacy violations, enabling users to maintain anonymity without waiting for new government regulations or a new technical standard. I conclude with an assessment of current trends in web privacy.

Privacy on the Web

In 1995, I created a demonstration web page called The Snooper to illustrate how much information a web site could gather on a user just from his or her having visited a single page. The results surprised me and many others. The user's email address, geographical location, computer

type, operating system, web browser and previous web site visited were all usually available. More recently, the HTTP "cookie" protocol has enabled advertisers to track a single user's browsing patterns over multiple web sites, maintaining a personal-interest profile. Combining this information with other publicly-accessible databases such as phone directories, marketing data, voter registration lists, etc. makes it possible for web sites to compile a significant amount of personal data on every visitor to their pages. All of this information-gathering may be accomplished without the user's authorization or awareness.

The Snooper employs several techniques, none requiring the use of "cookies," to gather its information about visitors. First, it analyzes the HTTP_USER_AGENT, REMOTE_HOST, and HTTP_REFERER variables, which almost all web browsers--including Netscape Navigator and Microsoft's Internet Explorer (IE)--provide to each site visited as part of the HTTP protocol. Here are typical settings for these variables:

```
HTTP_USER_AGENT = Mozilla/3.01Gold (X11; I; Linux 2.0.27 i586)

REMOTE_HOST = luck.sp.cs.cmu.edu

HTTP_REFERER =
http://altavista.digital.com/cgi-bin/query?what=web&q=AIDS+HIV+support+groups
```

HTTP_USER_AGENT reveals the user's browser software, which the remote web site could use to generate web pages specifically tailored to the browser's capabilities. However, both Netscape and IE also see fit to include the user's computer type and operating system as part of this variable. In the example shown above, Netscape ("Mozilla") is telling every web site I visit that I am running a Pentium with the Linux operating system.

The REMOTE_HOST variable reveals the Internet address of the computer making the request for a web page. Whether this information compromises the user's privacy depends on the type of Internet connection the user has. If the user is at a single-user workstation, as is typically the case in university environments, then the computer's identity may be the key to an enormous source of personal information. Using the Unix "finger" command, the Snooper script can uniquely identify the user's full name, email address, and often phone numbers in this case. At the opposite extreme, if the user is accessing the web via a large commercial provider such as America Online, or from behind a corporate firewall, then REMOTE_HOST may reveal simply that the user is an America Online member or an employee of that particular company. People who access the web via local Internet service providers (ISPs) reveal the identity of that ISP, which in turn reveals their geographic location. The Snooper script performs a "whois" lookup from the InterNIC database to find and report the physical address associated with the user's Internet host.

Finally, the HTTP_REFERER variable reveals the previous page visited by the user. For example, when the user performs an Internet search (with, say, Lycos or AltaVista) and then follows a link to reach a web site, that site is told exactly what search query the user had performed. It is possible to imagine cases where this information, provided without the user's consent or knowledge, compromises privacy.

The Snooper employs an additional trick to discover the user's email address: it includes an inlined image which forces the user's browser to request an *anonymous ftp* file transfer. Some browsers routinely provide the user's email address as part of the protocol for performing an ftp transfer; when this occurs, the Snooper script can record that email address. Netscape versions 1.0, 1.1, and

2.0beta all used this convention, thereby revealing the email address of every user to every web site they visited. Beginning with version 2.0, Netscape changed this behavior, and they awarded me a "Bugs Bounty" for pointing out the problem. New versions of Netscape and MSIE do not have this problem.

The Snooper also exploits another bug, present in Netscape versions through 2.0, which sends email from the user to any predesignated address without the user's knowledge. For example, just by visiting a page that exploited this bug, your computer could have emailed a death threat to `president@whitehouse.gov`; or, it might have simply sent an empty message to the web site owner. In this latter case, yet again, you would have revealed your email address. Although these egregious bugs have been fixed in newer versions, they demonstrate all too clearly that as writers of browser software add ever more features, they may also be introducing loopholes which can be exploited to compromise users' privacy.

A well-publicized example of a new browser feature compromising privacy is the so-called "magic cookie." Using cookies, a web site can tag each user with a unique identification number, which that user then presents, invisibly, for all future visits to that site. With the ability to recognize individual users each time they revisit a site, web sites can compile and accumulate profile information on their users over time. More ominously, cookies are allowed to be stored not only by the web sites you visit but also by the *images* displayed on web sites you visit--in particular, banner advertisements. Unbeknownst to most users, many of the Internet's ads reside on centralized ad servers run by agencies such as DoubleClick, Focalink, and Smartad. What this means is that the ad agency can, in principle, track a single user's browsing behavior over all the different sites which display that agency's ads. For example, as of this writing, DoubleClick manages the banner ads for AltaVista, U.S. News, Quicken Financial Network, and Travelocity. In principle, then, the agency could use cookies to build a single profile combining information about a user's web-searching, news-reading, financial and travel preferences. According to DoubleClick's privacy policy, they use the information thus collected for precision ad targeting but do not include the user's name or email address in the profile they build. Still, some find disturbing the notion of an advertising agency building a detailed profile of each user's browsing habits without the user's consent or awareness.

To summarize, although surfing the web feels anonymous, it is not. The technology underlying web browsing makes it possible for web sites to collect varying amounts of personal information about each user of their sites without consent. The TRUSTe Project, a joint effort by the Electronic Frontier Foundation and CommerceNet, proclaims a first principle of Internet commerce:

> Informed Consent is Necessary -- Consumers have the right to be informed about the privacy and security consequences of an online transaction BEFORE entering into one.

Current technology violates this principle. However, the Anonymizer provides a partial solution.

The Anonymizer

Technical Issues

The Anonymizer provides a technological means for preserving a user's privacy when surfing the web. The basic idea is very simple: set up a third-party web site (`http://www.anonymizer.com`)

to act as a middleman between the user and the site to be visited. When the user wants to view web pages at, say, the Apple Computer site, he does not ask his browser to establish a direct Internet connection to `http://www.apple.com`, but instead asks his browser to connect to `http://www.anonymizer.com:8080/www.apple.com`. The Anonymizer then makes the connection to `apple.com` without revealing any information about the user who requested the information, and finally forwards the information received from Apple to the user.

The basic principle of interposing a middleman server between user and web site is hardly novel. Indeed, the Internet firewalls used by most companies rely on "proxy servers," which use very similar technology to achieve their goal of eliminating direct connections between their employees and the outside net. The first version of the Anonymizer was based on the public-domain CERN proxy server, but with several modifications to preserve anonymity:

1. it does not forward the source IP address of the end-user;
2. it eliminates revealing information about the user's machine configuration from the "User-Agent" MIME header;
3. it eliminates the user's name from the "From" MIME header;
4. it eliminates the previously-visited site name from the "Referer" MIME header;
5. it does not forward the user's email address to serve as a password for FTP transactions;
6. it filters out Java applets and JavaScript scripts which may compromise anonymity;
7. it filters out all "magic cookies" which may compromise anonymity; and
8. it gives positive feedback to the user by displaying an Anonymizer header on the page and adding the word "[Anonymized]" to the page's title.

Furthermore, the Anonymizer provides an easy-to-use interface which allows users to bypass the cumbersome configuration procedure normally associated with using a proxy. This provides the following features:

1. users access the service simply with extended URLs, such as `http://www.anonymizer.com:8080/www.apple.com/`;
2. crucially, all embedded hypertext links in returned pages are automatically rewritten so that anonymity will be preserved when further links are followed;
3. users may freely intermix "anonymized" URLs with regular URLs during their browsing, so that the delays associated with the Anonymizer's extra redirection are incurred only when necessary.

The Anonymizer cannot guarantee its users perfect anonymity. One way in which anonymity can be violated is through the use of "helper applications," such as RealAudio, which go around the proxy by establishing their own direct net connections. Further, the technical standards underlying the Web are constantly in flux; changes to the HTML language can potentially create new routes around the Anonymizer's automatic link-rewriting mechanism. Nevertheless, in the vast majority of cases, users of the Anonymizer can feel secure that the sites they visit will learn only one thing about them: that they are users of the Anonymizer.

Practical Issues

The Anonymizer was originally developed on my home PC during the summer of 1995. Marc Ringuette helped me formulate the idea, and Darrell Kindred and Dayne Freitag contributed

valuable pieces of source code. I completed the implementation and put a test version of the system into operation at the Carnegie Mellon University School of Computer Science in the fall of 1995. During 1996, the software was licensed to C2Net, Inc. of Berkeley, CA, where it underwent public beta-testing. In 1997, I sold the software to Infonex, Inc. of La Mesa, CA. Both C2Net and Infonex are well-known for their commitment to Internet privacy issues. C2Net, owned by Sameer Parekh, is a leading provider of Internet cryptography software, and Infonex is co-owned by Lance Cottrell, who authored the Mixmaster anonymous remailer.

The identity of the new Anonymizer owners is important, because, unfortunately, trust must play a role when the Anonymizer is used. As a middleman server, the Anonymizer could in principle track its users' browsing patterns and make unscrupulous use of that information. In the case of the Anonymizer, Infonex's longstanding reputation as a privacy advocate is reassuring. The design of the Internet makes it technically very difficult to achieve anonymity without trusting an intermediary.

Also of practical concern is the Anonymizer's bandwidth requirement. As a publicly-accessible proxy server, the Anonymizer has been asked to handle hundreds of thousands of page requests daily for users around the world. Each request requires the Anonymizer to fetch, process, and forward a web page from elsewhere on the net. To pay for the hardware and network resources required to support the system, the owners of the Anonymizer are experimenting with a combination of advertising and premium subscription sales. As of July 1997, the free public Anonymizer service incorporates a one-minute delay in order to prevent severe overloading. The premium subscription service operates with a performance penalty of only a few seconds.

Future Directions

Because so many users have expressed anxiety about web privacy--as demonstrated not only by the popularity of the Anonymizer, but also quantitatively by a BCG/TRUSTe Internet Privacy study--the Internet's corporate powerhouses have proposed a new Open Profiling Standard (OPS). OPS, with the support of more than 100 companies including Netscape and Microsoft, aims to give users "personal control over the selective disclosure and sharing of personal information" (Lohr, 1997; "Proposal for an open profiling standard," 1997). If OPS is adopted, users will enter a personal information profile into their local web browsing software and then have the choice of which web sites they trust to access selected parts of that profile. From a privacy standpoint, this technology improves the user's control over released personal information when compared with cookie technology. However, it remains to be seen whether the type of information collected by "The Snooper"--which does not make use of cookies--will be at all curtailed by OPS. One important piece of information, the user's Internet host identity, cannot be hidden by OPS, because of the Internet's design. Only intermediary-based schemes such as the Anonymizer can protect that information.

The Anonymizer was the first, but is no longer the only, anonymizing proxy server available on the web. Two competing services are iproxy and the Lucent Personalized Web Assistant (LPWA). LPWA does not offer the Anonymizer's page-rewriting mechanism which enables users to easily change between anonymized and non-anonymized browsing. However, it does provide an additional feature: support for anonymous authentication and registration at web sites which provide personalized services.

The Anonymizer provides a practical solution to the problem of shielding one's identity from the end web sites that users visit. It does not solve the problem of shielding one's browsing patterns from the local system administrators at one's school, workplace, or Internet Service Provider. A planned future version of the Anonymizer will use encryption to implement this feature. Combining user-side encryption and site-side indirection will offer users the absolute maximum in privacy while they surf the web.

References

- Camp, L. J. (1997, February). Web security & privacy: An American perspective. ACM SIGCAS CEPC '97 (Computer Ethics: Philosophical Inquiry). Previous version presented as "Privacy on the Web", The Internet Society 1997 Symposium on Network & Distributed System Security, 10-11 February 1997, San Diego, CA.

- Lohr, S. (1997, June 12). Rare alliance on privacy for software. *New York Times*, p. C1.

- "Proposal for an open profiling standard." (1997, June 2). Document Version 1.0. Firefly Network Inc. Press Release. http://www.firefly.net/OPS/

Justin Boyan (Justin.Boyan@cs.cmu.edu) is a graduate student in Computer Science at Carnegie Mellon University. His doctoral research focuses on statistical algorithms for machine learning; his side interests in politics and privacy led him to author the "Anonymizer" software.