A reprint from
American Scientist
the magazine of Sigma Xi, The Scientific Research Society

Uniquely Me!

How much information does it take to single out one person among billions?

Brian Hayes

Suppose you fill out a survey online, with the assurance that your answers will remain anonymous. The questionnaire doesn't record your name and address, but it does ask for some demographic information: your date of birth, your zip code, and your gender. What are the chances you could be identified from those three facts alone? You can answer this question for yourself at the website http://aboutmyinfo.org, which was set up by Latanya Sweeney of Harvard University. In my case, the site reports that I am probably the only male born on December 10, 1949, living in zip code 02144. Thus three items of not-very-intimate information—gender, zip, birth date—reveal enough to pick me out of a crowd.

Ideas about identity, privacy, and anonymity are changing fast in this era of big data and social networks. At the deepest level, identity is all about the sense of self—the answer to the question "Who am I?" Each of us also has a biological identity (manifested in fingerprints, facial features, DNA sequences) and a legal identity (name, Social Security number, signature, and so on). Now we also have a data identity, defined by various combinations of traits that distinguish us from the rest of humanity. If you ask me to identify myself, I will not answer "M, 02144, 12/10/49"; and yet, by the combinatorics of uniqueness, I am that person as much as I am "Brian Hayes." Maybe more so: Dozens of people share my name.

Brian Hayes is senior writer for American Scientist. *Additional material related to the* Computing Science *column can be found online at http://bit-player.org. E-mail: brian@bit-player.org*

In the online world we have still more identities, most of them unknown even to ourselves. For example, I am my web browser history. The list of URLs I have visited in the past week or the past month is surely unique to me, just as my fingerprints are. I could even be identified by the list of fonts available to my web browser—and a few companies make use of such facts to track individuals as they wander from site to site across the web.

The Arithmetic of Uniqueness

When I first heard about Latanya Sweeney's demonstration that gender, zip code, and birth date are enough to identify many Americans, I found the result surprising, but the arithmetic is straightforward. For a back-of-the-envelope calculation, assume there are 300 million people in the United States, half male and half female, and that they are evenly distributed over 30,000 zip codes and 36,500 possible birth dates. (I am ignoring leap years and centenarians.) Each zip code has 5,000 male residents and 5,000 females. The question then becomes: If each of 5,000 people has a birth date chosen at random from 36,500 possibilities, how many will wind up with a date not shared by any other member of the group? The mathematically expected number is 4,360, or 87 percent.

The foregoing calculation is only a crude approximation. The real U.S. population is not distributed uniformly either by age or zip code. People in larger cohorts and more populous areas can more easily hide in the crowd. Philippe Golle of the Palo Alto Research Center has published an estimate of identifiability based on census data. He finds that the proportion of people with a unique combination of gender, zip code, and date of birth is a little over 60 percent.

Sweeney began her work on "re-identification" in the 1990s, when she was a graduate student at MIT. Her particular concern was the privacy of medical data. In 1997 she examined a batch of hospital documents released for statistical purposes and was able to identify the records of William Weld, a former governor of Massachusetts. The anonymized data listed each patient's gender, five-digit zip code, and date of birth, which Sweeney cross-linked with voter registration rolls. (Weld confirmed that the records were his.)

Partly in reaction to this incident, the Health Information Portability and Accountability Act (HIPAA) of 2003 established guidelines for guarding patient confidentiality. In general, aggregated medical data must not reveal exact dates of birth or precise locations.

Anonymous But Well Known

In a recent critique of Sweeney's re-identification work, Daniel C. Barth-Jones of Columbia University points out that a combination of attributes can't be proved unique without a "perfect population register," which lists the corresponding attributes of every person in the population. A perfect register is seldom available. Voter rolls are not even close to complete because not everyone votes. In the absence of a perfect register, an identification is a matter of probabilities—an assertion that coincidence is unlikely but not impossible.

The same argument applies to other identifying traits. I can't be certain that my fingerprints or my DNA are unique because I can't compare them with everyone else's. Nevertheless, such

biometric markers are used routinely in contexts where misidentification would have the gravest consequences. Of course the probability of uniqueness for fingerprints is thought to be very high—certainly higher than the 60 percent calculated for a combination of gender, zip code, and birth date. One hopes that no one will be sent to jail on the basis of a match to those three facts.

The standard of proof is quite different when the aim is preserving privacy rather than convicting an accused criminal. If you promise confidentiality to the subjects of a medical experiment, even a tentative identification represents a breach of trust.

Last year Sweeney and two colleagues published a follow-up study based on documents from the Personal Genome Project, where people voluntarily post their own genomic data for public access, annotated with whatever personal information they choose to disclose. Among 579 files that included gender, zip code, and birth date, Sweeney's group was able to match 130 to unique entries in voter lists; the Genome Project administrators confirmed that at least 121 of those names were correct.

In some contexts, matching unique data to a conventional identifier such as a name and address is beside the point. An Internet advertiser, for example, can make excellent use of a profile that reveals your interests and activities, even though the data are not linked to you by name. Indeed, the advertiser may prefer such "anonymous" data because there are fewer legal constraints on its collection and use.

History Sniffing

On the Internet, they say, nobody knows you're a dog. But everything else about you becomes marketing data for sale or trade.

Sharing information is what the Internet is all about, but most of us would like to retain some measure of control over the process. In particular, when you visit a website that doesn't require you to log in with a user name and password, you might think you could remain anonymous. But some sites go to extraordinary lengths to assign you a uniquely identifying profile.

One notorious technique is called history sniffing. Web browsers keep a list of visited URLs for the convenience of the user; the list is not supposed to be available to the websites you visit. But ingenious programmers have found ways to probe the list's content.

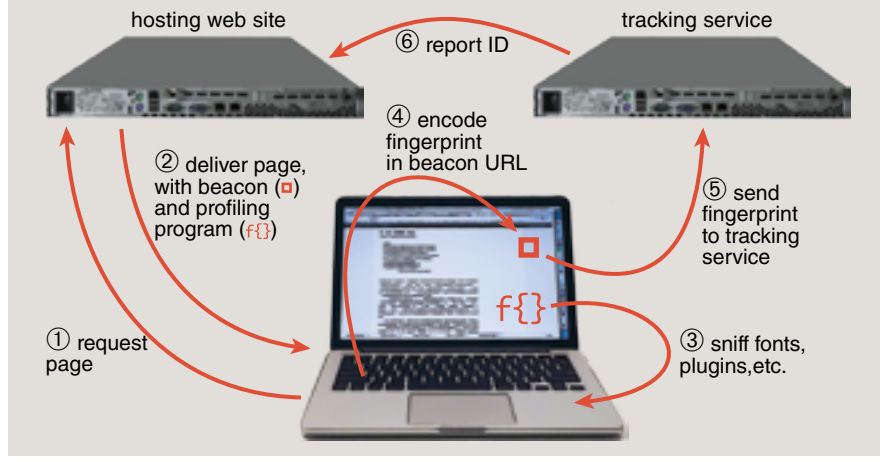
Browsers *do* offer a method to detect stylistic features of displayed information, such as the color of text. And visited links can be styled differently than unvisited ones. These facts set the scene for a privacy leak. An inquisitive website can include—hidden somewhere in the content it sends you—a list of links to various URLs. Also downloaded is a program (written in the JavaScript language) that checks the displayed color of each link. For every link that shows up as having been visited, the program sends a signal back to the web server.

This procedure does not answer the direct question, "What URLs are on your history list?" But it answers a series of yes-or-no questions of the form,"Have you recently visited site *X*?" Compiling a useful profile in this way might require asking about thousands of sites, which makes the technique grossly inefficient. But all the work of running the JavaScript program is done by *your* computer, not by the website's server. And the web user whose browsing habits are being recorded is generally unaware of what's going on; the long list of URLs is never actually displayed.

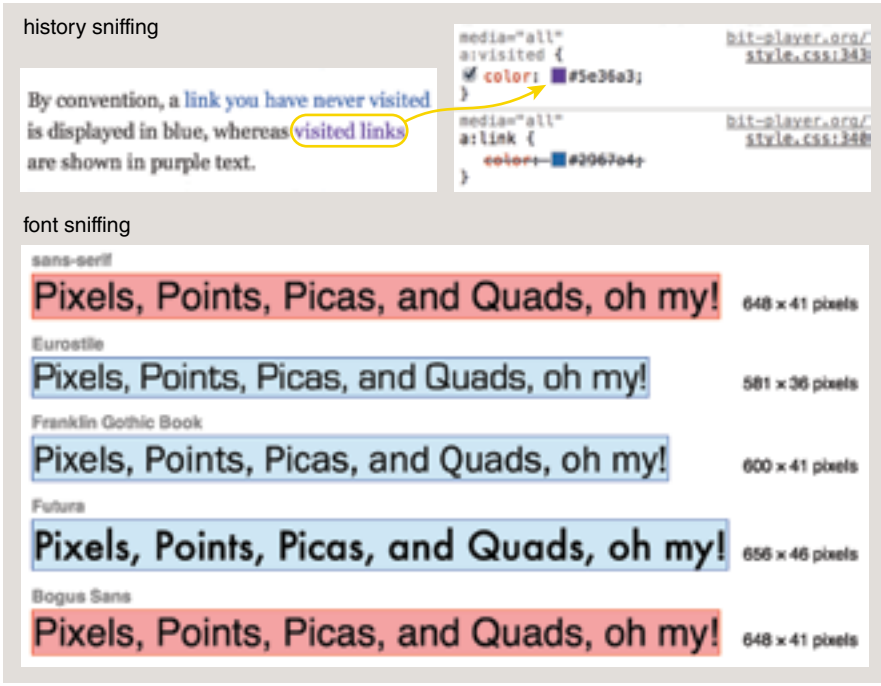
The operator of a website might be eager to peer into your history list for several reasons. For example, an online merchant might like to know if you have been shopping the competition. But even if the specific sites on the list are not of interest, the spectrum of yes-or-no responses can serve as an identifying fingerprint. What are the odds that your browsing history

Zip code, gender, and date of birth provide enough information to uniquely identify many Americans. Open bars show total population by age range in zip code 02144; solid bars indicate how many people in each age category can be expected to have a unique date of birth.

I am my browser history. The list of URLs I have visited is surely unique to me, just as my fingerprints are.



Fingerprinting software embedded in a web page gathers information about a user's browser, creating a uniquely identifying profile. After the software has done its work, it can embed the profile in the URL of a small image called a web beacon; when the browser later loads this image, it also sends the profile to a tracking service, which may share it with other sites.



Schemes for gathering information about a web browser are ingenious as well as devious. History sniffing detects differences in the formatting of visited and unvisited links. Font sniffing measures the dimensions of a string of text displayed with various typefaces. If a font doesn't exist (as with "Bogus Sans"), the text is displayed in a default font (in this case "sans-serif"). Checking a large number of links or fonts yields a unique "fingerprint" of the browser.

sets you apart from all others? Łukasz Olejnik of INRIA Grenoble and two colleagues collected history profiles from consenting volunteers. Out of 223,000 profiles in which they were able to detect at least four visited sites, 98 percent of the profiles were unique.

History sniffing has some defensible uses, but the potential for abuse was recognized early on, and recent versions of major browsers attempt to block history probes. Visited links are still rendered distinctively on the screen, but if a JavaScript program asks about that formatting, the browser lies, reporting that all links are unvisited.

In spite of these countermeasures, history sniffing has not disappeared. Last year a company called Dataium was accused of using history sniffing (among other techniques) to track the activities of automobile shoppers across 10,000 websites; in a negotiated settlement, Dataium agreed to abandon the practice. An earlier case against the advertising network Epic Marketplace reached a similar conclusion.

Meanwhile, other devious history-sniffing methods have come along. Instead of examining the format of a link, a program can measure the time needed to load an image from a site; a quick response to the request probably indicates that the image was already present in your browser's memory cache following a recent visit.

Font Sniffing

The history list is not the only part of a browser that a nosy website might try to sniff at. Peter Eckersley of the Electronic Frontier Foundation has cataloged a number of other browser properties that might also serve as identifiers. An intrusive program can enumerate the plug-ins or extensions installed in the browser, probe the list of fonts available for displaying text, or count the pixels on the computer's screen.

Are plug-ins, fonts, and other such attributes of a web browser likely to provide a uniquely identifying portrait? This might seem unlikely, in that computers ship with built-in fonts, and browsers come with a standard set of plug-ins, and many users never meddle in such technical arcana. Eckersley investigated the question by experiment. Among volunteers who visited a website set up to perform profiling, he found that almost 84 percent of browsers "had an instantaneously unique fingerprint." You can check your own browser configuration at https://panopticlick.eff.org. When I visited recently, the site reported: "Your browser fingerprint appears to be unique among the 3,760,699 tested so far."

One method of detecting fonts is similar to the trick for probing the history list. A website can request that text be displayed in a specific font; if the typeface is not available, the browser falls back to a default. The idea, then, is to ask for a sequence of characters to be rendered in many different fonts, and invoke a JavaScript function to measure the width and height of the resulting text. If the dimensions differ from those of the same character sequence in the default font, then the requested typeface must be installed on the user's computer and available to the browser. (As with history sniffing, all the formatting and measuring can be done out of sight, without actually displaying anything on the screen.)

Browser designers could take steps to prevent font profiling through JavaScript, but it's probably not worth the bother. There's an easier way to get font information from browsers that have an Adobe Flash plug-in (as most do): The Flash scripting language includes a command to list all installed fonts.

A group of investigators at the Catholic University of Leuven have surveyed a million websites to see how many are exploiting intrusive technologies such as font sniffing. The reassuring news is that only a tiny fraction of the sites—perhaps one in a thousand—seem to be engaging in the most devious practices. On the other hand, a few of those sites are apparently large and popular ones.

Browser profiling is not always done for nefarious purposes. A bank might use a browser fingerprint to trigger extra security precautions when a customer logs in from an unfamiliar location. But even when the aims are legitimate, companies tend to be secretive about the practice. One prominent website that appears to engage in browser fingerprinting is the Skype telephone service. Skype's 5,000-word privacy statement does not clearly disclose that fact.

The tracking methods I have described here are especially sneaky, but they are hardly the only threats to personal privacy on the Internet. Most tracking relies on "cookies" (text that a website can store in your browser) and "beacons" (links to images or other objects that reveal your arrival on a web page). The more elaborate sniffing

methods may be aimed primarily at those who block cookies and beacons.

33 Bits of Information

Fifteen years ago, when the public Internet was still young, a Silicon Valley executive dismissed concerns about privacy in online life. "You have zero privacy anyway," he said. "Get over it." The remark was jarring at the time, but it seems that many of us *have* gotten over it—or else given in to it.

For a major segment of the population, the urgent concern is not privacy but sharing: We tweet, we link in, we update our status. Although these communications are meant for a select audience, most people understand that everything they post on a social network is also visible to the operators of that network, and perhaps to others. It's a bargain they make willingly: A fifth of humanity is on Facebook. But no one willingly submits to font sniffing and other surreptitious profiling schemes.

Plugging such privacy leaks is hard. The root of the problem is that each of us really is unique, not only in deep matters of body and mind but even in our most trivial attributes, such as the cruft we've squirreled away over the years in dusty corners of a computer disk. In a world where every tiny idiosyncrasy can be cataloged and filed away in milliseconds, it's all too easy to compile a unique fingerprint. Just 33 bits of information is enough to single out any one person from the world population of 7.1 billion.

In some contexts, thoughtful attention to counting those bits has helped to draw a curtain of discretion over personal data. The HIPAA regulations for medical data are an example, and the Census Bureau has similar policies. For example, population breakdowns by race and sex are not released for the smallest geographic divisions, and various kinds of random noise are added to some tabulations. The study of such measures—asking how best to protect individual identity without impairing the research value of the statistics—has grown into a thriving minidiscipline called differential privacy.

Perhaps some variant of the same approach can be made to work for everyday life online. Website designers would still get enough information about the browser environment to present information effectively, but they wouldn't get 33 bits.

Bibliography

Acar, G., et al. 2013. FPDetective: Dusting the web for fingerprinters. In *Proceedings of the 20th ACM Conference on Computer and Communications Security*, pp. 1129–1140.

Barth-Jones, D. C. 2012 preprint. The "re-identification" of governor William Weld's medical information: A critical re-examination of health data identification risks and privacy protections, then and now. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2076397.

Eckersley, P. 2010. How unique is your web browser? In *Proceedings of the 10th Privacy Enhancing Technologies Symposium*, pp. 1–17.

Golle, P. 2006. Revisiting the uniqueness of simple demographics in the U.S. population. In *Proceedings of the Fifth ACM Workshop on Privacy in Electronic Society*, pp. 77–80.

Nikiforakis, N., et al. 2013. Cookieless monster: Exploring the ecosystem of web-based device fingerprinting. In *Proceedings of the 2013 IEEE Symposium on Security and Privacy*, pp. 541–555.

Olejnik, Ł., C. Castelluccia, and A. Janc. 2013. On the uniqueness of web browsing history patterns. *Annals of Telecommunications* doi:10.1007/s12243-013-0392-5.

Sweeney, L. 2000 preprint. Simple demographics often identify people uniquely. Data privacy working paper 3, Carnegie Mellon University. http://dataprivacylab.org/projects/identifiability/paper1.pdf.

Sweeney, L., A. Abu, and J. Winn. 2013 preprint. Identifying participants in the Personal Genome Project by name. http://privacytools.seas.harvard.edu/publications/identifying-participants-personal-genome-project-name.