

COMPUTERWORLD Networking & Internet

 Print Article  Close Window

What Google knows about you

Google may know more about you than your mother does. Got a problem with that?

Robert L. Mitchell

May 11, 2009 ([Computerworld](#)) "Google knows more about you than your mother."

Kevin Bankston, senior staff attorney at the Electronic Frontier Foundation, recently made that statement to this reporter. A few years ago, it might have sounded far-fetched. But if you're one of the growing number of people who are using more and more products in Google's ever-expanding stable (at last count, I was using a dozen), you might wonder if Bankston isn't onto something.

It's easy to understand why privacy advocates and policymakers are sounding alarms about [online privacy](#) in general -- and singling out Google in particular. If you use Google's search engine, Google knows what you searched for as well as your activity on partner Web sites that use its ad services. If you use the [Chrome browser](#), it may know every Web site you've typed into the address bar, or "Omnibox."

It may have all of your e-mail ([Gmail](#)), your appointments (Google Calendar) and even your last known location ([Google Latitude](#)). It may know what you're watching (YouTube) and whom you are calling. It may have transcripts of your telephone messages ([Google Voice](#)).

It may hold your photos in Picasa Web Albums, which includes face-recognition technology that can automatically identify you and your friends in new photos. And through [Google Books](#), it may know what books you've read, what you annotated and how long you spent reading.

Technically, of course, Google doesn't know anything about you. But it stores tremendous amounts of data about you and your activities on its servers, from the content you create to the searches you perform, the Web sites you visit and the ads you click.

Google, says Bankston, "is expecting consumers to trust it with the closest thing to a printout of their brain that has ever existed."

How Google uses personal information is guided by three "bedrock principles," says Peter Fleischer, the company's global privacy counsel. "We don't sell it. We don't collect it without permission. We don't use it to serve ads without permission." But what constitutes "personal information" has not been universally agreed upon.

“

[Google] is expecting consumers to trust it with the closest thing to a printout of their brain that has ever existed.

Kevin Bankston, senior staff attorney, EFF

Google isn't the only company to follow this business model. "Online tools really aren't free. We pay for them

with micropayments of personal information," says Greg Conti, a professor at the U.S. Military Academy at West Point and author of the book *Googling Security: How Much Does Google Know About You?* But Google may have the biggest collection of data about individuals, the content they create and what they do online.

It is the breathtaking scope of data under Google's control, generated by an expanding list of products and services, that has put the company at the center of the online privacy debate. According to Pam Dixon, executive director at the World Privacy Forum, "No company has ever had this much consumer data" -- an assertion that Google disputes.

Opacity vs. transparency

Critics say Google has been too vague in explaining how it uses the data it collects, how it shares information among its services and with its advertisers, how it protects that data from litigators and government investigators, and how long it retains that data before deleting or "anonymizing" it so that it can't be tracked back to individual users.

"Because of Google's opacity as to how it is using that data, and a lack of fundamental information rights [that] users have, [privacy] becomes a very thorny question," says Dixon.

Privacy policy opacity isn't limited to Google. It's so prevalent, in fact, that the Federal Trade Commission warned the industry in February that online businesses will face increased regulation unless they produce privacy statements that explain in a "clear, concise, consumer-friendly and prominent" way what data the companies collect, how they use it and how users can opt out ([download PDF](#)).

Google, however, contends that the concerns about opacity and the scope of data it collects are overblown. "I do push back on this notion that what we have is a greater privacy risk to users," says Mike Yang, product counsel in Google's legal department. Google, he says, gives users plenty of transparency and control. "There's this notion that an account has a lot more information than is visible to you, but that tends not to be the case. In most of the products, the information we have about you is visible to you within the service."

In fact, though, the data Google stores about you falls into two buckets: user-generated content, which you control and which is associated with your account; and server log data, which is associated with one or more browser cookie IDs stored on your computer. Server log data is not visible to you and is not considered to be personally identifiable information.

These logs contain details of how you interact with Google's various services. They include Web page requests (the date, the time and what was requested), query history, IP address, one or more cookie IDs that uniquely identify your browser, and other metadata. Google declined to provide more detail on its server log architecture, other than to say that the company does not maintain a single, unified set of server logs for all of its services.

Google says it won't provide visibility into search query logs and other server log data because that data is always associated with a physical computer's browser or IP address, not the individual or his Google account name. Google contends that opening that data up would create more privacy issues than it would solve. "If we made that transparent, you would be able to see your wife's searches. It's always difficult to strike that right balance," Yang says.



Google has been collecting all of this information over time about people and they said they would not be using that data.

Nicole Ozer, technology and civil liberties policy director, ACLU of Northern California

You do have more control than ever before. Google says it removes user-generated content within 14 days for many products, but that period can be longer (it's 60 days for Gmail). For retention policies that fall "outside of reasonable user expectations or industry practice," Google says it posts notices either [in its privacy policy](#) or in the individual products themselves.

You can control the ads that are served up, either by adding or removing interest categories stored in [Google's Ads Preferences Manager](#) or by opting out of Google's Doubleclick cookie, which links the data Google has stored about you to your browser in order to deliver [targeted advertising](#). For more information, see "[6 ways to protect your privacy on Google](#)."

Shuman Ghosemajumder, business product manager for trust and safety at Google, says users have nothing to worry about. All of Google's applications run on separate servers and are not federated in any way. "They exist in individual repositories, except for our raw logs," he says. But some information is shared in certain circumstances, and Google's privacy policies are designed to leave the company plenty of wiggle room to innovate.

Yang points to [Google Health](#) as an example. If you are exchanging messages with your doctor, you might want those messages to appear in Gmail or have an appointment automatically appear in Google Calendar,

he says.

Google is hoping that [what it lacks in privacy policy clarity](#), it can make up for in the transparency of its services.

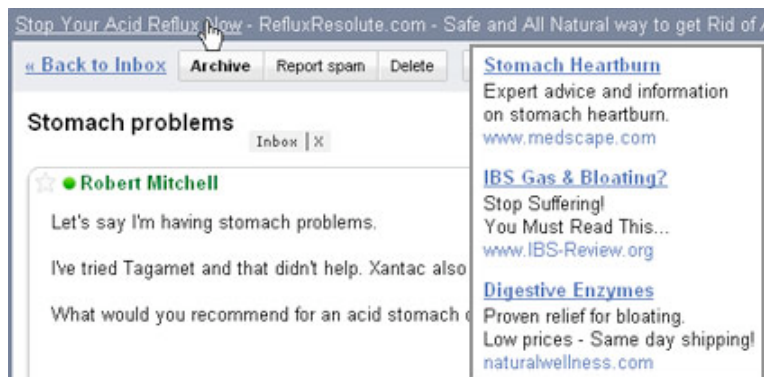
But Dixon, who follows medical privacy issues, contends that they aren't transparent enough. Medical records, once transferred to Google Health, aren't protected by HIPAA or by the rules of doctor-patient confidentiality. Google states that it has no plans to use Google Health for advertising. But by sharing data across services, the company is blurring the lines, Dixon says.

If you have a health problem and you use Google Health, research the disease using Google's search engine, use Gmail to communicate with your doctor, and link appointment details to Google Calendar, and your last location in Latitude was a medical clinic, Dixon asks, "What does the advertiser get to know about you? What about law enforcement? Or a civil litigant? Where are the facts? I don't have them, and that bothers me."

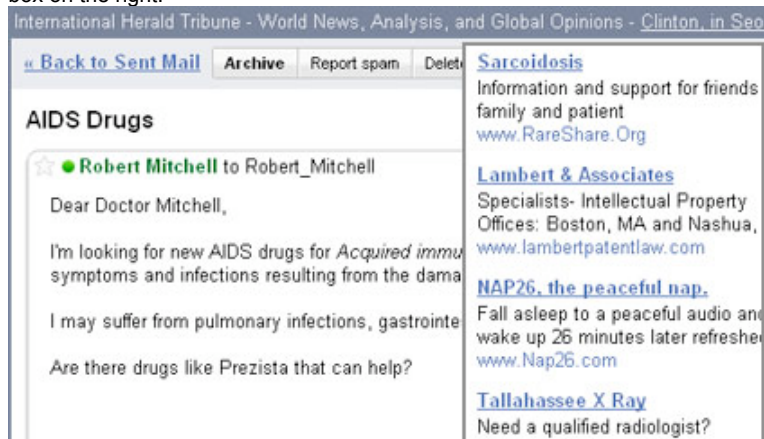
Change in behavior

Google's recent decision to change gears and mine what it knows about you to better target advertisements has also raised concerns.

Until recently, Google placed ads based on "contextual targeting" -- derived from the subject of a search or a keyword in a Gmail message you were reading, for example. To avoid creeping people out with ads targeting sensitive subjects, it avoids the topics of race, religion, sexual orientation, health, political or trade union affiliation, and some sensitive financial categories, as illustrated in the screenshots below.



Google uses the contextual information in a Gmail message to target the ads it serves, as shown in the inset box on the right.



In this case, the context of the e-mail is too sensitive for targeted advertising, so nontargeted ads are served.

With the information at its disposal, Google could pull together in-depth profiles of its users and launch highly targeted ads based on who you are (your user profile) and your activity history on the Web. The latter is a controversial practice known as behavioral advertising. Until recently, Google rejected the technique.

Then, on March 11, Susan Wojcicki, vice president of product management, announced in a post on Google's official blog that the company was taking a step in that direction. With the [launch of "interest-based advertising,"](#) Google is beginning to target ads based not just on context but on the Web pages you previously viewed.

That Web page history will come from a log associated with the cookie ID. However, since that ID links not to

a unique user but to a unique browser, you may end up viewing ads for Web pages visited by your spouse or others who share your machine. In a bizarre Catch-22, advertisers will be able to target ads at you based on logs that Google says it cannot make available to you -- for privacy reasons.

Ghosemajumder acknowledges that the situation isn't perfect. "In some cases there is [transparency], and in some cases there isn't," he admits. But he says Google is "trying to come up with more ways to offer transparency."

Privacy advocates fear that interest-based advertising is just the first step toward more highly targeted advertising that draws upon everything Google knows about you. "This is a major issue, because Google has been collecting all of this information over time about people and they said they would not be using that data," says Nicole Ozer, technology and civil liberties policy director at ACLU of Northern California.

But privacy advocates say Google is also doing some things right, such as launching its online [Privacy Center](#) and providing [additional controls](#) for some of its services.

Google is not acting alone in moving toward behavioral advertising. It is simply joining many other companies that are pursuing this practice. Mike Zaneis, vice president of public policy at the Internet Advertising Bureau, acknowledges that highly targeted advertising can be creepy. But, he says, "creepiness is not in and of itself a consumer harm."

The practice is unlikely to change unless users respond by abandoning services that use the techniques. But he argues that they won't because highly targeted ads are of more interest to users than nontargeted "spam ads."

Concerns have also been raised about Google's ability to secure user content internally. Google has had a few small incidents, such as when it allowed some [Google Docs users' documents to be shared](#) with users who did not have permission to view them. But that incident, which affected less than 1% of users, pales in comparison to security fiascoes at Google's competitors, such as [AOL's release of search log data](#) from 650,000 users in 2006.



In some cases there is [transparency], and in some cases there isn't.

Shuman Ghosemajumder, business product manager for trust and safety, Google

Ghosemajumder says the privacy of user data is tightly controlled. "We have all kinds of measures to ensure that third parties can't get access to users' private data, and we have internal controls to ensure that you can't get access to data in a given Google service if you're not part of the team," he says.

How anonymous?

Bowing to pressure, Google has made other concessions as well.

Google doesn't delete server log data, but it has agreed to anonymize it after a period of time so that the logs can't be associated with a specific cookie ID or IP address. After initially agreeing in 2007 to [anonymize users' IP addresses](#) and other data in its server logs after 18 months, it announced last September that it was [shortening that period to nine months](#) for all data except for cookies, which will still be anonymized after 18 months. "All of our services are subject to those anonymization policies," says Ghosemajumder.

Critics complain that Google doesn't go far enough in how it anonymizes personally identifiable data. For example, Google zeroes out the last 8 bits of the 32-bit IP address. That narrows your identity down to a group of 256 machines in a specific geographic area. Companies with their own block of IP addresses also may be concerned about this scheme, since activity can easily be associated with the organization's identity, if not with an individual. Even anonymized data can be personally identifiable when combined with other data, privacy advocates say.

Sensing an opportunity, and facing similar criticisms, competitors have tried to go Google one better. Rather than anonymizing IP addresses, Microsoft deletes them after 18 months and has proposed that the industry anonymize all search logs after six months. Yahoo anonymizes search queries and other log data after three months, and the Ixquick search engine doesn't store users' IP addresses at all.

Perhaps the biggest concern for privacy advocates is how the treasure trove of data Google has about you [might end up in the wrong hands](#). It is, says Bankston, a wealth of detailed, sensitive information that provides "one-stop shopping for [government investigators](#), litigators and others who want to know what you've been doing."

Privacy laws provide little protection in this regard. Most policies -- including Google's -- don't provide an explicit guarantee that the company will notify you if your data has been requested through a court order or subpoena. "The legal protections accorded to data stored with companies [and] the data they collect about you is very unclear," says Bankston.

The industry still relies on the [Electronic Communications Privacy Act of 1986](#), a 22-year-old privacy law that even the government has argued doesn't apply to online data. "Google has yet to state a specific position on whether and how that law protects your search logs," Bankston says.

Privacy groups want Google to reveal just how many requests it receives from litigators and law enforcement and how it responds to those requests, but the company, like its competitors, does not release that information.

Google declined to elaborate on why it's not more forthcoming in this regard, but deputy general counsel Nicole Wong did say that it complies with legal requests "narrowly, appropriately and in accordance with the law."

In at least one high-profile case, Google has taken strong measures to protect the privacy of its subscribers. When Viacom issued a subpoena for the viewing records of Google's YouTube subscribers, it fought the subpoena and turned over only anonymized data that it says can't be traced back to individual users.

But privacy watchers question what happens to the thousands of requests for individual records in less prominent cases. Google's response: "Our overarching principle is we want to notify users," says product counsel Mike Yang.



For a company with so much data, they have a responsibility to be innovative, proactive and pro-consumer in the area of privacy.

Pam Dixon, executive director, World Privacy Forum

The ACLU's Ozer thinks Google should collect less data and store it for shorter periods of time. That's one of her suggestions in a 44-page [privacy and business primer](#) for Web 2.0 companies published by ACLU of Northern California.

With lawmakers focused on the economy, privacy groups say it's unlikely that laws will change anytime soon. But they -- and regulators -- are pressuring Google to provide leadership and set the example. "For a company with so much data, they have a responsibility to be innovative, proactive and pro-consumer in the area of privacy," Dixon argues.

Parting with a certain amount of personal information is part of the bargain you strike when you sign up for free Web-based software and services. "People should have visibility into what information is being collected and how it will be used, and they should get to choose what they share and control who can access it," says Fleischer.

But the tension between your desire for privacy and Google's need for flexibility in handling your data is likely to be an ongoing dance. "Google's business is to make money from the information it gathers from its users. It will always be a give-and-take," says Bankston.

Eventually, updated privacy laws -- and the choices users make -- will delineate what is acceptable and what is not. "Our business model depends completely on [user trust](#)," says Yang. Part of Google's challenge is to build that trust without severely restricting the business's ability to innovate.

Google has done a good job on the trust side, Bankston says. He just wants to see it give users more transparency and control. "We don't want Google to stop innovating," he says. "We just want the law to keep up so that this data is safe."

Next: [6 ways to protect your privacy on Google](#)