The Cost of Free Web Tools

Edward Sobiesk and Gregory Conti
US Military Academy

Your company is secretly planning a merger—they know about it. Your CEO's child has recently come down with a rare illness—they know about it. A key staff member in your organization surfs for pornography at night using his home computer—they know about it. Who is this Orwellian entity that has so much information about your organization? It's not the government, but rather the companies that provide free Internet services such as search, mapping, and email. In other words, it's Google, Yahoo, Microsoft, AOL, MapQuest, and other major Web-based service providers.

It's doubtful that these companies are currently receiving daily briefings about the internal operations, strategies, and people in your organization. However, their servers almost certainly retain sufficient information (including every word your employees type into search boxes) to let them know such things if they wished—especially if they were to focus on an organization or a given set of individuals, perhaps via an IP address range or registered user accounts. Free Web-based services aren't really free: users pay for them with micropayments of information that add up to a very significant sum.

Unknown disclosures

Most users assume that their use of Internet services is implicitly private and anonymous, so it can be quite eye-opening to find out how much about ourselves and our companies we reveal by the seemingly innocuous words we use to search, the maps we view, and the other "free" services we use on the Internet. The Internet has become one of the most central aspects of our world, and we react to both the mundane and important events in our personal and professional lives by turning to it. Unfortunately, these events, great or small, will continue to exist for an indeterminately long time period on the service providers' servers.

This information's continued persistence, combined with the increasing ability to tie it to real-world users, is a risk that most individuals and organizations don't realize they assume. The amount of information accumulated about our online activities and behaviors has reached unprecedented magnitude. This information is one of the most precious corporate assets for the companies that amass it, and its great value guarantees that many will covet it. When asked why he robbed banks, the famous early 20th century US bank robber Willie Sutton allegedly responded, "Because that's where the money is." In the 21st century, information is the currency of the day, and the data centers of Google, Yahoo, Microsoft, and AOL are the banks.

The sensitivity of retained data—and the risk associated with assuming that interactions are private exchanges between users and online services—is perhaps best evidenced by the August 2006 AOL data set disclosure in which the company inadvertently made available to the public the search terms for approximately 20 million Web searches by 658,000 AOL users.[1] This landmark incident confirmed what security researchers and privacy experts had suspected:[2,3] that data retained from the use of Web-based tools presents a significant privacy issue. Although the 20 million AOL searches weren't released with AOL usernames or IP addresses attached, the data set did include unique numeric identifiers in place of each username. Using this "anonymized" identifier and its associated search queries, *The New York Times* reporters Michael Barbaro and Tom Zeller quickly demonstrated the ease with which someone could use the search queries to identify actual people who had conducted the searches.[4] Web sites such as www.aolstalker.com and www.aolpsycho.com have amplified this process by allowing their visitors to collaboratively analyze, tag, and in some cases identify users from the AOL data set. When we recently visited the AOL Stalker site, for example, we noted that user #672368's queries had been viewed 30,813 times, and he or she had been rated as "entertaining." This person's queries range from the commonplace to the highly sensitive, as do many of our own, addressing religion, pregnancy, shopping, and abortion clinics in Charlotte, NC. If you choose to review the AOL data set yourself, sites like AOL Stalker will not only track what you search for within the data set, they'll sometimes also let other users of the site know exactly what you're looking for.

The initial AOL incident and subsequent exploitive Web sites fostered a spurt of articles and blog dialogues[5,6] further highlighting the significant privacy issues at stake here. The media coverage briefly brought the issue of information disclosure and data retention to the forefront of public awareness, but it appears that the AOL disclosure's long-term impact on public consciousness has been minimal. Although most privacy experts still see the incident as a watershed event, a recent survey of college undergraduates and a corresponding pilot survey of nontechnical middle-aged users revealed that 83 and 84 percent, respectively, of them had no familiarity with the incident.[7] For the typical user, the disclosure essentially didn't occur.

Prior to the AOL data set's release, the realization of how much information we divulge on the Internet had been slowly moving into the public consciousness. One of the forerunners in identifying this phenomenon was John Battelle who coined the phrase "database of intentions" to describe the information gathered by recording all our searches:[8]

> "The Database of Intentions is simply this: the aggregate results of every search ever entered, every result list ever tendered, and every path taken as a result. It lives in many places, but three or four places in particular—AOL, Google, MSN, Yahoo—hold a massive amount of this data."

Battelle and a few other contemporary writers[3] have made strong cases for how much information about ourselves, our organizations, and our society is revealed through our use of Web-based services.

Reasons for data retention

It's important to realize that online service companies have very significant reasons for accumulating user data.

Business models

Providers of free Web-based applications aren't simply offering their tools as a public service. However altruistic they might be in some regards, these companies have legal obligations to their shareholders to make profits. Although various business models exist for advertising in connection with "free" services, the consistent bottom line is that Web-based companies depend on being able to convince advertisers that it's worth their money to have their ads presented on Web pages and emails. This model earns the service provider only a few cents for each click-through, but when you consider that Google handles more than 100 million searches a day,[9] it's easy to see that customized, targeted advertising is a multibillion-dollar industry.

Innovation and service improvement

The primary way that providers improve their services is through research, and meaningful research almost certainly requires data. Every Web-based service provider's goal is to become a household name such as Google, Microsoft, or Yahoo. In this intense battleground, those currently on top expend immense effort to maintain their positions, as evidenced by Google's world-class research facilities, top-tier scientists, and mammoth processing assets, all of which are equivalent to nation-state-level resources.[10]

Legal requirements

Although still growing in significance, legal requirements could soon be a third major reason that Web-based companies retain extensive amounts of data. Online service companies and Internet service providers (ISPs) tend to resist such measures, citing administrative and technical overhead as well as cost. Another often unstated reason for their resistance is public opinion. Online service companies and ISPs depend on their users' trust, and any degradation of it represents an attack on their corporate bottom line. That said, lawmakers around the world are increasingly seizing on the Web's ability to retain data trails as a key tool in law enforcement. Perhaps the potential for increased legal requests of archived search data was a driving factor behind Google's recent decision to remove the associated IP addresses from retained user queries after 18 to 24 months.[11]

Balancing innovation, profit, and privacy

Although a complete solution to the challenges of data disclosure and data retention currently lies beyond reach, we believe that self-monitoring tools, anonymous browsing capabilities, and open dialogue, all of which would raise public awareness, represent excellent starting points.

Development of easy-to-use self-monitoring tools, such as an enhanced browser history function that lets users review all the information (including search terms) they've disclosed, is an important first step. Integrated into browsers and employed on corporate networks, such tools would raise awareness about how much information is being divulged and could encourage individuals and organizations to intelligently self-regulate their disclosures. Ideally, the online companies would be at the forefront of these initiatives and would help create tools that covered not just search but all outbound Web-based information flows.

Anonymous browsing technologies will continue to improve and increase in availability. However, users still have to trust some organization (probably several) with their information. These might include an ISP, online service provider, motherboard manufacturer, operating system developer, and Web anonymizer company. A key requirement in dealing with this issue is for users to view information disclosure as a personal

responsibility rather than just someone else's problem.

An open dialogue between the Web-based companies, policymakers, organizations representing user privacy rights, and individual end users is a critical step toward working through these challenges. The work of the Electronic Frontier Foundation (www.eff.org), the Electronic Privacy Information Center (www.epic.org), and Lauren Weinstein[3] are excellent examples of how the brightest privacy advocates are engaging this issue so far.

Free Web-based services aren't really free: users pay for them with micropayments of information that add up to a very significant sum.

Clearly, no single solution will solve all the challenges—not least because future tools and services will raise new issues that need further attention.

In today's era of customized advertising, we see a trend toward automated mining and correlating of our data disclosures, which has a host of ethical implications. Free email accounts, such as Gmail, frequently use machine processors to examine email content and add relevant advertising. But what constitutes going too far? For instance, one of our friends noticed, as he was involved in an email exchange about a recently deceased relative, that his free Web-based email account was including advertisements for bereavement counseling with the messages.

Potential privacy concerns also arise with emerging free services such as Wink.com and Spock.com, which crawl the Web accumulating publicly available material, including that located on social networking sites, to put together profiles on people. *The Wall Street Journal* recently ran an eloquent discussion of this issue in an article entitled, "The Story of Your Life, Now Available Online."[12] Such sites reinforce how information that seems innocuous in isolation can be revealing when consolidated in an organized, meaningful manner.

Because improving Web-based services, both from performance and business perspectives, will likely require greater information about the user, privacy issues related to search technology and the aggregation of data through free Web-based services will continue to grow. As the Web's functionality evolves to facilitate both human users and software agents acting on their behalf (a concept that's a cornerstone of the Semantic Web), further information disclosure and data retention will be required. Anticipating and working through these challenges now, before they become crises, is a goal from which all parties would benefit. □

Acknowledgments

The views expressed in this article are those of the authors and do not reflect the official policy or position of the US Military Academy, the Department of the Army, the Department of Defense, or the US government.

References

1. R. Singel, "AOL's Search Gaffe and You," *Wired News Online*, 11 Aug. 2006; www.wired.com/news/politics/privacy/0,71579-0.html.
2. G. Conti, "Googling: I'm Feeling (un)Lucky," presented at DEFCON 14, 4 Aug. 2006; www.defcon.org/html/defcon-14/dc-14-speakers.html#Conti.
3. L. Weinstein, "An Open Letter to Google: Concepts for a Google Privacy Initiative," 9 May 2006; www.vortex.com/google-privacy-initiative.
4. M. Barbaro and T. Zeller, "A Face Is Exposed for AOL Searcher No. 4417749," *The New York Times*, 9 Aug. 2006; www.nytimes.com/2006/08/09/technology/09aol.html?ex=1312776000&en=f6f61949c6da4d38&ei=5090.
5. M. Arrington, "AOL Proudly Releases Massive Amounts of Private Data," *TechCrunch*, 6 Aug. 2006; www.techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data.
6. "AOL Releases Search Logs of 657,427 Users," *Slashdot.org*, 7 Aug. 2006; http://yro.slashdot.org/article.pl?sid=06/08/07/2022244.
7. G. Conti and E. Sobiesk, "An Honest Man Has Nothing to Fear: User Perceptions on Web-Based Information Disclosure," *Proc. Symp. Usable Privacy and Security*, 2007; to be published.
8. J. Battelle, *The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture*, Portfolio, 2005
9. G. Gilder, "The Information Factories," *Wired*, vol. 14, no. 10, 2006, pp. 178–202; www.wired.com/wired/archive/14.10/cloudware.html.
10. G. Conti, "Googling Considered Harmful," *Proc. New Security Paradigms Workshop*, 2006; www.nspw.org/2006/.
11. S. Diaz, "Google to Tighten Privacy," *The Washington Post*, 15 Mar. 2007.
12. J. Vascellaro, "The Story of Your Life, Now Available Online," *The Wall Street J.*, 14 Mar. 2007, p. D1.

Edward Sobiesk *is an assistant professor in the Department of Electrical Engineering and Computer Science at the US Military Academy. He has a PhD in computer and information sciences from the University of Minnesota. Contact him at edward.sobiesk@usma.edu.*

Gregory Conti *is an assistant professor in the Department of Electrical Engineering and Computer Science at the US Military Academy. He has a PhD in computer science from Georgia Tech. Contact him at gregory-conti@usma.edu.*