When Spiders Bite: The Use, Misuse, and Unintended Consequences of "Silent Information"

Thomas P. Keenan

Professor, Faculty of Environmental Design, University of Calgary,
Calgary, Alberta Canada
keenan@ucalgary.ca

Abstract. Spiders are the workhorses of the Internet, silently (and almost invisibly) traversing the online world, 24 hours a day, looking for information that may be of interest to someone. It is being archived, organized, and sold, usually without the knowledge or consent of the subject of the information. Serious consequences are starting to appear, such as the withdrawal of three candidates from the October 2008 Canadian Federal election because of previous online indiscretions. While these were intentional if mis-guided postings, information made available without our consent can have equally devastating effects. Advances in artificial intelligence, as well as the increasing tendency to post more and more information, such as videos, will make the gathering, aggregation, and republishing of this "silent information" an increasingly important issue that must be addressed from the technical, social, ethical and legal perspectives, and sooner rather than later.

Keywords: Privacy, identity, profiling, data mis-use, tagging.

1 The Elephant in the Room

Well intentioned privacy experts, such as Canada's Privacy Commissioner, spend a great deal of ongoing effort [1] discussing the rules that should govern the collection and use of personally identifiable information. Despite cross-cultural differences, there is general agreement on the duty of companies and governments to handle personal information with great care.

Canada's *Personal Information Protection and Electronic Documents Act,* (PIPEDA) which received Royal Assent on April 13, 2000, requires that "an organization may collect, use or disclose personal information only for purposes that a reasonable person would consider are appropriate in the circumstances."

European countries are even more advanced in regulating the use of personal data, stemming in part from historical factors such as the Holocaust, "when the Nazis used public and church records to identify Jews to be rounded up and sent to concentration camps" [2] and manifested in national laws inspired by the landmark The European Union Directive on Data Protection of 1995 [3].

Even in the United States, which has a reputation of being less concerned about corporate invasion of privacy, and relatively more concerned about such action by

V. Matyáš et al. (Eds.): The Future of Identity, IFIP AICT 298, pp. 83–91, 2009.

governments, [2] credit bureaus are highly regulated and required to disclose information they hold on a person upon proper request. Many jurisdictions, notably the state of California, have enacted laws in the last five years mandating that breaches of private information be disclosed in writing. All of these are good and useful policies, but they generally refer to information collection and use that a person already knows about.

They are therefore missing the proverbial "elephant in the room."

In real life, vast amounts of personally identifiable information are being harvested, informally, in a variety of ways, in many jurisdictions, with no real consent of the subject. In many cases the subject of the information is unaware of the collection, and certainly not fully informed as to the ultimate destination of the information. This paper will confine itself to information that winds up on the Internet, partly because that medium has become the dominant way in which such information is shared, and also because, frankly, there is no way to really know what information is being collected and shared behind closed doors, though the author's previous paper [4] hints at the extent of government snooping projects such as CARNIVORE.

That paper, as well as an excellent study by Jones and Soltren, [5] demonstrated that social networking sites such as MySpace, Facebook and Nexopia are treasure troves of information about people that, with little difficulty, can be tied back to them. The privacy policies of such sites generally confer ownership of all content to the site operator, and provide little opportunity to effectively retract information, beyond simply deleting it and hoping for the best. Technical factors, such as the ease with which a digital photo can be copied off such sites, make any idea of "recalling your information" completely infeasible. You can virtually assume that if you post it, it will be copied by someone or archived somewhere.

The dominant automated technology for trolling the Internet for information is a robotic computer program called a "web crawler," or more commonly a "spider." Because computing power, storage, and bandwidth have become so inexpensive, it is now feasible for these agents to continually traverse the Internet, collecting whatever information they are directed to amass, copying it from webpages and other sources onto a company's own computers. Just as living spiders can inflict a painful or even fatal bite, the information gathered by computer spiders can cause harm by revealing personal or corporate data that was not intended to be shared. The damage is further complicated by the near impossibility of regaining control of that information once it has been harvested by a spider.

In addition to being easy to collect, information has become relatively easy to process into actionable intelligence. Artificial intelligence programs for extracting useful information and patterns from data have names like Adaptive Fuzzy Feature Map, (American Heuristics Corporation), PolyAnalyst 6.0. (Megaputer Intelligence Inc), and FactSpotter (Xerox). There are also companies that offer data mining as a professional service.

These existences of these programs and services provide a clue to the motivation of those who set spiders loose on the Internet. Just as biological spiders search for food and bite when threatened; these companies and individuals see economic advantage in collecting, organizing, using and selling the data that has been casually left around by others.

2 What Is "Silent Information"?

There is no word that comes to mind (at least in the English language) to precisely describe personal information that is available to others without the subject's explicit knowledge and/or consent. The closest would probably be Clarke's work [6] on "profiling" but that relates to the different problem of making inferences about someone, while silent information is already linked to a specific person.

As just one provocative example, USA/Israel-based Zoom Information Inc. (www.zoominfo.com) collects, without obtaining consent, a variety of information from mentions in newspaper articles, publicly posted presentations, etc. Just as Google's spiders prowl the Internet, looking for information to index, this company's robotic probes seek information that can be attached to an individual's profile. As of October 15, 2008, they claim to have information on over 44 million people and more than 4 million companies. Much of it is incomplete and incorrect. Most people do not even realize they are listed there. Yet ZoomInfo stays in business because some of their information is unique and so valuable in the business context (for example, senior executives and their contact information) that people will pay $99 US/month for access to it.

One could argue that posting a PowerPoint presentation that shows my job title as "Professor" is giving tacit permission for someone to put my name and that title into a database and sell access to it. However, I was also, to my chagrin, listed on ZoomInfo as Chairman of the Board of a US defense contractor, based on an erroneous name-based inference by the company's software.

Sometimes, the results of a ZoomInfo search are bizarre and even hilarious. The current Prime Minister of Canada, Stephen Harper, was listed for a period of time on this site with the job title "Reluctant Leader" and his company shown as "the Conservative party" simply because those terms appeared in a newspaper article about him. To be fair, there is a mechanism to "claim your profile" and correct errors, but the Honorable PM has not yet done so as of this writing.

Another source of potential embarrassment is the website DiplomacyMonitor.com, which archives press releases by the many countries for 90 days. Sometimes governments have been known to pull press releases off their official websites, but they remain readable at this site. There are also "whistleblower" sites such as www.wikileaks.org, which says it specializes in "source documents were classified, confidential or censored at the time of release."

During the Canadian federal election campaign, which culminated in the October 2008 vote, several candidates were forced to withdraw when embarrassing facts about them surfaced on the Internet.

Justin West, who was running for the New Democratic Party (NDP) in a riding in British Columbia, withdrew from the race after images of him swimming in the nude, which were over a decade old, were found on the Internet. Conservative Chris Reid was forced out because of evidence that he had made extremist comments on a personal blog. NDP hopeful Dana Larsen had to step down when a video of him smoking marijuana was found online. Past indiscretions simply never go away in the digital world.

More poignant and personal examples arise from companies such as ChoicePoint, a LexisNexis company which sells "comprehensive credentialing, background screening, authentication, direct marketing and public records services to businesses and nonprofit

organizations." It goes far beyond credit information, which in the US is regulated under the Fair Credit Reporting Act. For example, this company gathers data by sending people into courthouses to copy out court records.

One woman (in a personal communication with the author) said she had trouble obtaining credit because of this company's refusal to include the full details of a Small Claims Court appearance she made (in which she was actually the plaintiff). The company has admitted to other blunders involving private information, at least one on a massive scale. In a Form 8-K regulatory filing with the US Securities Exchange Commission dated March 4, 2005, the company reported that "based on information currently available, we estimate that approximately 145,000 consumers from 50 states and other territories may have had their personal information improperly accessed." A Nigerian national named Olatunji Oluwatosin pleaded "no contest" to identify theft charges arising from this data breach, and was sentenced to ten years in prison. ChoicePoint was fined $10M US.

2.1 A Difficult but Necessary Definition

For the purposes of this paper, "silent information" is defined as **"**person-linked information which is deliberately collected and distributed without the subject's explicit understanding and consent to the full range of its ultimate use**."**

The "person" could be a real human being, a company, a login name, a Second Life avatar, etc. This linking may be "explicit" (personally identifiable information is included) or "implicit" (it is possible to deduce the subject's identity). This would therefore exclude aggregated or demographic data, which is frequently being sold, except where the sample size is so small as to effectively disclose the identity of the subject.

Another dimension relates to the authoritativeness of the linking of data item A to person B. ZoomInfo, as discussed above, often mis-attributes data to people who have similar names, so it would generally be less authoritative than, for example, an official driver's license database.

Clearly the subject of "profiling" is closely related to "silent information." Even if it is not personally identifiable, the collection of data on a group of people can be used to take measures that may affect them. So, for example, Google provides aggregated data on the movies that are viewed by college students. This can be used by movie distribution companies to plan advertising campaigns to maximize their profits. While not, strictly speaking, an invasion of privacy, the net effect is some manipulation of behavior imposed upon a group that is, quite probably, unaware of how this is happening.

"Location-based marketing", in which for example SMS messages are sent to a user as he or she walks by a store, takes this one step further, leading people to wonder, "How did they know I would enjoy a latte right now?" The key point is not the intrusiveness of such a technique, but the fact that its operation is not transparent to the person being targeted.

2.2 What Is Consent?

The most challenging part of this definition is probably the phrase "without the subject's explicit understanding and consent." Precisely what does that mean? Is having accepted a posted privacy policy sufficient evidence of understanding and consent? Probably not.

Such policies are usually ignored. Khosrow-Pour's study [7] of over 261 US business graduate students, found high awareness (77.4% had "seen a privacy policy statement") but also low interest (54.4% said they had not "read a privacy policy statement"). Another study [5] of Facebook-using students at MIT, Harvard, NYU and the University of Oklahoma found 91% had not read the site's Terms of Service and 89% had "never read the privacy policy."

In addition, having plowed through a statement crafted in legal language does not imply understanding the full implications. For example, what if you have agreed to post something on website A and it is then harvested and posted on website B without your knowledge? The author was called by someone who objected that a memorial tribute she had written at a funeral home's website had been posted on another site "without my consent". She had basically lost control of this writing once it was posted online, and it didn't "feel right" to her.

The "rules of engagement" when it comes to the re-use of information posted online are often hazy because it is impossible to predict all the possible ways in which information could be re-purposed, now and in the future. In fact, information that may seem to have no or very slight privacy implications may well become very intrusive in the future. Consider, for example, the blood samples routinely taken from babies when they are born. In some places, these have been archived for decades and, now, with modern technology, they could suddenly become a treasure trove of DNA information.

2.3 What Is "Deliberate vs. Accidental Disclosure?"

There are certainly breaches that happen without the consent of the site operator. Ample illustration of this came in the August 2008 "Facebook virus" crisis, in which someone was able to send messages to Facebook users that appeared to be from their "Friends" on the system. The goal was to get victims to download and execute the "codecsetup.exe" file which installs the GAMPASS virus. The unauthorized use of the profile photos of "Friends" conferred credibility to the attack, and was a triumph of social engineering against Facebook users, who have been characterized as "notoriously naïve when it comes to security awareness" [8].

A 2007 study by Sophos PLC [9] revealed that 41% of Facebook users contacted at random divulged personal information to a fictitious person, created for the study. At the same time a Facebook spokesperson estimated that less than 20% of users have changed their privacy settings from the default.

Site operators like Facebook can be victim of their own internal errors, even in the absence of malicious outsiders. Sophos reported in July 2008 that personal data on 80 million Facebook users was compromised because "a security slip-up by the website during the process of a public beta test of its new design for members' profiles left birth date information exposed"[10]. Facebook quickly fixed the problem.

One might argue that at least Facebook users took some voluntary action to post their information, such as real birth date, even if their intention for its use was not properly respected. An even clearer case of accidental disclosure happened at Columbia University. On June 10, 2008, the Vice President responsible for Student Auxiliary & Business Services wrote individual letters [11] to a large group of people, noting that "one archival database file containing the housing information of approximately 5,000 current and former undergraduate students was found on a Google-hosted website" and

that “your name and Social Security number were included in the file”. While these students were probably aware that this information was in the hands of their university, they certainly never expected it to be found on the public Internet.

Accidents like this will happen, and the laws and policies to deal with them are still evolving, both in government legislation and in court cases. The author [12] suggested a mechanism which would provide monetary compensation to victims of identity theft if it could be linked back to gross negligence on the part of a company that held information on them, and in fact that is exactly how the TJX data breach is being handled, with up to $30 US in compensation being offered as compensation to most self-declared victims. (see www.TJXsettlement.com.)

2.4 Technical Issues Affecting Silent Information

It should be noted that website operators, Internet service providers, etc. may need to make use of user-posted information for purely technical reasons such as backups and system optimization, or to comply with lawful requests from appropriate authorities, including the removal of inappropriate content. Most jurisdictions, and certainly Canada in its PIPEDA act [1] make allowances for this.

One of the simplest forms of “silent information” is the website that sent you to another place on the Internet. While the designers of the original Internet Protocol suite probably never anticipated the extent of interest in your navigation history, it has become a “hot topic” because of its value in e-commerce settings, i.e. to pay the site that drives traffic to a commercial site.

There is no universal mechanism for keeping track of who really was visited an open website (one that does not require account/password type authentication) beyond recording the IP address and the referrer link. This mechanism is laid out in RFC2616 (July 1999) updated by RFC 2817 (May 2000). However, IP addresses are of limited utility and can be spoofed by knowledgeable users. Therefore, website operators turn to other mechanisms.

Cookies (strings of data left on a computer by a website to facilitate tracking on subsequent visits) have been identified as potential privacy violation and explicitly addressed, e.g. in the 2002 European Union telecommunication privacy directive [13] which requires the user be informed of the attempt to store a cookie and given the option to refuse it. Outside of Europe, cookies are still routinely set, though most browsers can be configured to refuse or challenge them.

Web bugs (usually small e.g. 1x1 pixel images, which require downloading from a server) are another tracking technique that can log whether web pages are being read and if so by what IP address. This is useful to spammers, e.g. for identifying “live” email addresses, and for other marketing purposes. Some email clients are counteracting this technique by asking explicitly for permission to download images.

3 Legal, Ethical, Policy and Social Issues

Non-technical policy questions abound in this area. Who owns information that has been harvested online? What rights do people have to get it corrected? How does this play out in a cross-border situation? What is the legal status of deliberately planted

"dis-information"? They are just starting to be addressed in the courts. In July, 2008 a UK judge awarded 22,000 GBP to a man who was the victim of "Facebook libel" because of a fraudulent and malicious profile created by others [14]. Lawyers are even being told to analyze their clients' online presence and to review any blogs and emails before their opponents can raise them in court [15].

In a case that is still before the Canadian and US courts, a person has posted a photograph of the leader of a Canadian organization with the comment "This person should be killed." There is a significant difference between the laws in those two countries regarding whether or not this is "protected speech" (e.g. under the First Amendment to the US Constitution), "hate speech", or an incitement to commit a criminal act.

Another fascinating development is the move to link an individual's DNA information to other personal information held in databases. DNA is collected by law in many jurisdictions from those convicted or even accused of crimes. It is also surrendered voluntarily for purposes such as paternity testing. Google has invested in at least two companies (Navigenics and 23andMe) that work in this area. DNA provides a non-refutable, highly authoritative validation of identity. Once a link is made between a person's DNA and online persona, the concept of online anonymity becomes essentially moot.

Even without DNA linking, it has been shown, for example by George Danezis at the FIDIS/IFIP Summer School in Brno in September, 2008, that assumptions about gaining true anonymity by the use of "anonymizer" programs such as Tor (available at www.torproject.org) may well be unfounded. In addition, information that is now effectively anonymous, or of little interest, can easily be saved and processed in the future, with greater computing power and superior algorithms, leading to a retroactive breach of privacy.

Should we have expectations of privacy in the online world? Do we? At least one study [16] suggests that bloggers have abandoned any notion of privacy. The move to "open source" content provides further evidence of a trend towards freer dissemination of information. Just as philosophers and lawyers needed to define a community standard for obscenity 50 years ago, perhaps we will need an understanding about what is "reasonable privacy" in the near future.

4 Information Tagging – At Least a Partial Solution

The inability of some of the largest media companies in the world to prevent piracy of their movies and music gives a clue as to how difficult it would be to actually keep track of the spread and possible misuse of one's personal information. Just as someone can sneak a video camera into a movie theatre or concert, if one can display information on a screen, it's possible to capture and re-enter it manually, obliterating any attempt to control it. But it's still worth trying.

Let's consider a very restricted problem – tracking where your "Facebook profile photo" may have migrated online. You could "watermark" it visibly, noting that you do not authorize further distribution. There is also steganography software that will allow you to put invisible codes into the image to help in identifying it. Europol already administers [17] a database of the checksums of images used in child pornography investigations to assist in subsequent cases.

But how would you know where to look for your photo? We may be getting close to a solution for that. It turns out that several companies, notably Google [18] and Idée Inc**.** [19] are working on "visual search engines" to help content owners track images by visual identification, even if the image has been altered, e.g. by Photoshop. The latter company's TinEye project is now in a public beta trial, but as of August, 15, 2008, only accesses 701 million images so a visual search using my Facebook profile photo didn't show any proliferation. However, Hillary Clinton's photo popped up as being copied in 22 different places as strange as www.thisisbigbrother.com.

5 The Future

Should personal information be explicitly "tagged" for its acceptable use? How would we even anticipate those uses? Several commentators have noted that the very act of tagging information, e.g. "this is my driver's license number, please do not mis-use it" invites the abuse that it is trying to prevent. Even saying "this is confidential" provides a temptation to some people to snoop. Indeed, the very foundation of "whistle-blower sites" like www.wikileaks.org relates to people's very natural curiosity about things that they are not supposed to see.

One way in which tagging may have value is in the legal enforcement of rights to information such as our photographs. Although the holders to music and movie copyrights have had a difficult time enforcing their rights, it is definitely true that putting a label on information indicating that it is not simply "there for the taking" provides useful legal support in acting against those who have appropriated the information. Still, as noted in a previous paper, [4] the most popular places people are posting information generally assert that they own the information posted there. So the creator, of, say, a Facebook profile photo, might not actually be able to do much if that image was mis-used.

Perhaps we should just follow the precautionary principle and never make something available online which might come back to hurt us? Doing that would greatly limit our "online presence" and perhaps still fail to protect us effectively. It seems clear that some combination of technical, legal, ethical and educational measures will be needed to preserve the public's confidence in both personal privacy and freedom to communicate.

References

1. Privacy Commissioner of Canada, Privacy Act Reform, http://www.privcom.gc.ca/legislation/pa/pa_reform_e.asp (accessed August 14, 2008)
2. Sullivan, B.: 'La difference' is stark in EU, U.S. privacy laws, http://www.msnbc.msn.com/id/15221111/ (accessed August 14, 2008)
3. http://www.cdt.org/privacy/eudirective/EU_Directive_.html (accessed August 15, 2008)
4. Keenan, T.: On the Internet Things Never Go Away Completely. In: Fischer-Hübner, et al. (eds.) The Future of Identity in the Information Society, pp. 37–50. Springer, Boston (2008)

5. Jones, H., Soltren, H.J.: Facebook: Threats to Privacy, http://groups.csail.mit.edu/mac/classes/6.805/student-papers/fall05-papers/facebook.pdf (accessed August 15, 2008)
6. Clarke, R.: Profiling: A Hidden Challenge to the Regulation of Data Surveillance, http://www.anu.edu.au/people/Roger.Clarke/DV/PaperProfiling.html (accessed August 15, 2008)
7. Khosrow-Pour, M. (ed.): Technologies For Commerce and Services Online, p. 103. IGI Global (2008)
8. Arrington, M.: Elaborate Facebook Worm Virus Spreading, http://www.techcrunch.com/2008/08/07/elaborate-facebook-worm-virus-spreading/ (accessed August 15, 2008)
9. Sophos, PLC, http://www.sophos.com/pressoffice/news/articles/2007/08/facebook.html (accessed August 14, 2008)
10. Sophos PLC, http://www.sophos.com/pressoffice/news/articles/2008/07/facebook-birthday.html (accessed August 15, 2008)
11. Wright, S.: Personal letter dated June 10, 2008, sent to Columbia University students
12. Keenan, T.: When Disclosure Becomes Spam – The Apparent Failure of Well Intentioned Privacy Policies and Legislation and How to Fix Them. In: Proceedings, 49th Annual Conference of the Western Social Science Association, Calgary, AB (2007)
13. http://eurlex.europa.eu/smartapi/cgi/sga_doc?smartapi!celexapi!prod!CELEXnumdoc&lg=en&numdoc=32002L0058&model=guichett (accessed August 15, 2008)
14. Richards, J.: Fake Facebook profile victim awarded 22,000 GPB, http://technology.timesonline.co.uk/tol/news/tech_and_web/article4389538.ece (accessed August 15, 2008)
15. Menzies, K.B.: Perils and possibilities of online social networks: has your client been networking on Facebook or MySpace? If so, you need to know about it - because your opponent will. In Trial 44(7), p. 58
16. Viégas, F.B.: Bloggers' Expectations of Privacy and Accountability: An Initial Survey, Media Laboratory, Massachusetts Institute of Technology, http://jcmc.indiana.edu/vol10/issue3/viegas.html (accessed August 15, 2008)
17. Leyden, J.: http://www.theregister.co.uk/2003/04/14/us_gov_builds_huge_child/ (accessed August 15, 2008)
18. Beet TV, http://searchengineland.com/080715-084529.php (accessed August 15, 2008)
19. Personal communication, Leila Boujnane, CEO, Idée Inc., Banff, Alberta (June 7, 2008)