

Privacy vs. Profiling on the Internet

Csaba Kiraly

Department of Information and Communication Technologies,
University of Trento

kiraly@dit.unitn.it

It is just getting common knowledge that private data entered on web pages is not safe at all. What is less known is that on the Internet, even if using encryption, simply by communicating, the communication pattern (services we connect to, amount and timing of communication) reveals a lot about ourselves. Attackers, or even the Internet Service Provider itself can use Statistical Traffic Analysis techniques to gather information, to profile users for marketing or for other (even malicious) reasons.

To face this threat, privacy preserving technologies have been developed since the early eighties, providing anonymous communications and masquerading traffic patterns.

Are the privacy enhancing technologies ready to be used by the masses?

Is it legal to use such technologies in a world where strong encryption was banned for private use even some years ago? Where retention of communication data (for several years) is gaining ground in legislation?

Are we becoming safer or more of a suspect by using such technologies?

Problems of Privacy and Profiling

In the summer of 2006 AOL (America Online), one of the largest Internet Service Providers in the United States, was making their database of search query logs available to researchers. The database contained about 19 million Internet search queries of their 658.000 users, made during a period of three months.

Of course it was well known to AOL that, in order to protect the privacy of their users, such data should be sanitized before release, so they have anonymized it replacing user identifiers with random numbers. Unfortunately, they have made two errors that lead to a case picked up by US media, and we can say, to a privacy scandal. First, instead of just sending the data to the researchers, it was released to the Internet¹. This would of course not have been a problem in case the data have been perfectly (if such thing exists) sanitized. Second, they have only replaced the user IDs, but the search queries were still organized per user with their entire text. The example below shows some queries of a user².

User ID	Keywords	Rank	Date
4417749	mature living	1	2006-05-28 14:55:19
4417749	apartments 28806 area	10	2006-05-28 14:53:52
4417749	apartments 28806 area	9	2006-05-28 14:53:52
4417749	apartments in west asheville north carolina		2006-05-28 14:52:43
4417749	apartments in asheville north carolina		2006-05-28 14:52:02
4417749	yerba mate		2006-05-28 12:08:32
4417749	tea for good health	4	2006-05-28 12:06:55
4417749	immune	2	2006-05-28 12:02:59
4417749	sinus infection	2	2006-05-28 11:54:22
4417749	termites	2	2006-05-28 11:53:05

It was rather easy even for journalists (and not professionals specialized in such data collection) to find the real identity of people based on their queries, which sometimes even contained their real names or their home address.³

Once a person is identified, how much the totality of these queries tells about him? Well, it could be information like one's travel plans, including options considered. But it can also be related to one's friends, interests, health problems, etc.

The incident of releasing this data was not in itself harmful. We could even say that a search query, typed in, could evidently leave a trace, and people, especially after a number of related media scandals could be aware of this. However it was pointing out how enormous is the amount of data logged day-by-day, and how easily it can get into the wrong hands even without intentional disclosure.

What is less known is that communication leaves other kinds of traces as well, even if the communication itself was encrypted. The most obvious one of these traces is the address of

1 See e.g. *Chronicle of AOL Search Query Log Release Incident* available at http://sifaka.cs.uiuc.edu/xshen/aol_querylog.html (September 8, 2007)

2 The example is taken from a copy of the original database, still available at many places, e.g. from <http://data.aolsearchlogs.com/search/Index.cgi>

3 See the article of *the New York Times: A Face Is Exposed for AOL Searcher No. 4417749* available at http://www.eff.org/Privacy/AOL/exhibit_d.pdf

the two communicating parties, which might then be used for example to find out who made a specific, anti-regime description on Wikipedia⁴. It could also be used the other way around, starting from the user, and looking at the list of services he was accessing on the Internet, which nowadays, when on-line shopping took off, can reveal enormous amounts of information about the person.

It also starts to get a common practice of Internet Service Providers (ISP) to try to block traffic related to some kinds of user activities, typically peer-to-peer traffic (used for sharing files, which might violate copyrights). To achieve this blocking, traffic is analyzed and grouped into classes. This classification, however, can also be used for profiling users and their activities. We might ask why an ISP (who has the only role to provide communications services) should know how much and when we read our mail, we chat or call our friend, whether we like to share files and who we are sharing it with, etc.

Recent works have also shown that by looking at the traffic pattern (the order, size and timing of IP packets sent by/to the user's computer), even more can be done. It has been proved⁵ that the traffic pattern could reveal the website visited even if the communication is encrypted and the end-point is anonymized. Web pages can be characterized by a traffic pattern fingerprint, extracted by statistical methods from a number of traffic patterns captured during the download of that specific page. Later, looking at a traffic pattern generated by a user, it is possible to determine whether the contacted web site was e.g. Amazon, or CNN. It is easy to imagine how the same technology can be used to identify other sites, where the sole interest of the user is very sensitive and personal information. It can also be envisioned (even if not yet proved) that the same method can be used to identify actions taken on a specific site, e.g. whether someone was just looking around in a web-shop or something was bought.

Protection by Legislation or by Technical Means?

There is an ongoing debate about whether (and to what extent) traces of personal communications should be collected for reasons of public safety, a goal sometimes privileged over privacy. Besides its use in surveillance, such personal data also provably holds an enormous business value in the age of personalized marketing.

In some countries, laws are created to protect privacy and limit the cases where data could be logged and the time it could be kept. In others, very similar laws are prepared to oblige the collection of personal data for at least a number of years in order to perfect surveillance. European countries created legislation (parts of it harmonized throughout the EU) that treats privacy as a fundamental personal right. This right is in clear contrast with public and business interest, which can take precedence in some well defined, but (more or less clearly) bounded cases. The US adapts a model where personal data is seen more as property of the person, with the consequence of the data having a recognized value and being tradable.⁶

In such an environment, one might ask what a person can do in order to preserve his own privacy on the Internet. Besides expressing preferences through effecting legislation changes (definitely not an easy task), or trying to force existing personal information rights (not easy

4 Actually, the technique was already used to find "anonymous" Wikipedia edits that were made from IP addresses belonging to "interesting" organizations, such as governmental agencies, churches or multinational companies. See the article of the Wired magazine: *See Who's Editing Wikipedia - Diebold, the CIA, a Campaign*, available at http://www.wired.com/politics/onlinerights/news/2007/08/wiki_tracker, or the database itself at <http://wikiscanner.virgil.gr/>

5 See M. Liberatore, B. N. Levine, "Inferring the Source of Encrypted HTTP Connections", CCS2006, October 2006;

G. D. Bissias, M. Liberatore, D. Jensen, B. N. Levine, "Privacy Vulnerabilities in Encrypted HTTP Streams", PET 2005, Cavtat, Croatia, May 30-June 1, 2005

6 For an overview of differences, see Priscilla M. Regan, *Safe harbours or free frontiers? Privacy and Transborder Data Flow*, Journal of Social Issues, vol. 59, n.2, 2003.

as well), one could try to protect personal information himself. Gary T. Marx presents "eleven behavioral techniques of neutralization intended to subvert the collection of personal information"⁷. The technical means available to protect privacy on the Internet mostly fall in the *blocking/masking moves* categories according to Marx's classification.

Internet related Privacy Enhancing Technologies have been researched and developed since the early eighties and are available today for use: some only in research laboratories, while others are on the market as free or pay products. They sanitize information at different levels, providing anonymity, pseudonymity, or hiding the content and context of the communication. An interesting aspect of some of these technologies, namely, anonymous routing technologies, is that successful implementations rely on the communities running and using them. In these systems, anonymity exists only to the extent of being a non identified one of all the users.

Is There a Need for Privacy Protecting Technologies?

How much people are aware of potential privacy issues on the Internet? Whether they care at all? Are they more concerned about governmental or commercial organizations? How much would they sacrifice (whether we are speaking of money, services, quality of services, or a sensation of living in a monitored and therefore secure environment) in order to have their privacy? Whether they know about, use, or would use technologies aimed at preserving privacy? Are we becoming safer or more of a suspect by using such technologies?

As far as awareness is concerned, what can clearly be seen is some kind of media awareness in the United States. The AOL search query release was growing into a scandal quickly, with posts from many of the major newspapers and magazines, including The New York Times, Wall Street Journal, Washington Post, CBS News, to name just a few. This is reinforced by a law which obliges organizations to notify their users in case of a loss of personal data. Surprisingly, in Europe it seems hard to find similar examples. Whether it is due to the lack of cases, to the lack of a similar law or simply to the lack of interest, is not clear.

As an indicator of the need for privacy protection, we might also look at the price of some of the products offering (at least in their claims) privacy of communications. It is interesting to see how these products appeared in the last five years, as targeted marketing and spam messages gain ground. Their price now is at the level of 10-20 euros per year, not something how one would value an important fundamental right. Does this mean that privacy protection is not that important? I wouldn't say, rather, we should ask the question which other fundamental rights are we paying for this directly: we are not used to pay for them directly, as they are taken for granted.

Another indicating factor is download and usage statistics of products enhancing privacy. While most of the products were developed for research purposes or are cumbersome to install, some try to go further. One of these notable exceptions is the Freedom Network, which was building up and selling a service of anonymous communications starting from 1998. It was ceasing operation in 2001 due to the un-sustainability of their business model; as one of the co-founders stated,⁸ "Initially we got incredible response for the premium services, but we knew we were dealing with early adopters. But soon we saw the transfer into the mass market just didn't carry over. The subscription rates really plunged."

On the other hand, Tor, a free anonymizing service, is growing slowly. The difficulty of Tor is that it can only work if some of the users take on the role of providing so called exit

7 See Marx, Gary T., *A tack in the shoe: neutralizing and resisting the new surveillance*, Journal of Social Issues, vol. 59, n.2, 2003.

8 See the related article of The Register: *Zero-Knowledge bags anonymity service*, available at http://www.theregister.co.uk/2001/10/05/zeroknowledge_bags_anonymity_service/

nodes: computers that are easy to identify and where seemingly all the anonym traffic is coming from. Owners of these nodes frequently receive queries or requests to shut down their operation from governmental entities, even in democratic countries such as Germany or the United States. However, it seems that the argument of providing services to citizens of countries where contents available on the Internet would otherwise be censored, is strong enough to let them continue their operation.

Surveying Privacy Protection on the Internet

Turow and Hennessy address some of the previously posed questions in their recent survey "from a perspective of environmental risk and the public's trust of institutions".⁹ They point out that, just as with environmental risks, the specialty of privacy related dangers is that "they are not visible to the general population". The individual's opinion about privacy issues is formed based on the "belief in their existence", which is determined mainly by reports of dueling experts and by media attention, contrasted with their own (and usually missing) experiences. The telephone interview survey conducted in the US shows interesting results, but it does not address differences that are deeply founded in cultural and historical differences.

Regan¹⁰ presents another telephone interview based survey concentrating on consumer marketing, privacy legislation and trust in institutions of the private sphere. The survey, commissioned by IBM, provides an international comparison between the United States, Germany, and the United Kingdom. It shows that about 80% of the respondents "believe that they have lost control over both how information is collected and how it is used by companies". High level of pessimism is shown regarding the possibility of protecting privacy in the computer age as well. The survey also demonstrates some differences at national level, however, I think there are much more differences in attitudes and legislations that need further study.

Without evidence, intuition says that a comparative study carried out in Europe and the United States would reveal fundamental differences in the tradeoffs people accept when sharing their private data or when they expose themselves to surveillance. Confronting situation in the UK (where CCTV based surveillance is on one side largely accepted, on the other side it has already grounded privacy related debates) to other European countries could yield interesting results as well.

Another interesting cultural difference could be studied by confronting European post-communist countries with post-western societies. As Majtenyi writes¹¹, contrary to western constitutions, "personal data is guaranteed almost without exception by the constitutions of post-communist countries". In these countries, protection of personal data "conveyed the barely concealed political message that it was possible, indeed necessary, for the individual to counter the omnipotent state".

There are also fundamental differences in how the use of PET technologies is viewed. While Canada is on the forefront of research in PET technologies and government representatives actively participate in related scientific conferences, the German government seeks ways to forbid the use of some of these technologies.

The US has a long standing history of restricting the use of security technologies. After the Second World War, US government export policy has restricted the strength of the cryptographic algorithms available to the private sphere as well as for export, for national security considerations. Limitations prevailed till the late nineties, and it was assumed (although not

9 See Turrow Joseph and Hennessy Michael, *Internet privacy and institutional trust: insights from a national survey*, New media and Society, vol. 9, n. 900, 2007.

10 See Regan Priscilla M, *Safe harbours or free frontiers? Privacy and Transborder Data Flow*, Journal of Social Issues, vol. 59, n.2, 2003.

11 Majtenyj Laszlo, *Ensuring data protection in East-ecentral Europe*, Social Research, vol. 69, n.1, 2002.

proved) that encryption allowed for the private sphere was limited at a level which could be deciphered using technologies available to US governmental agencies, but not to commercial organizations, neither to foreign countries. However, these limitations proved to be unsustainable as decrypting capability of computers started to increase with exponential speed, leaving the information in this "ideal" state of being protected from some but decryptable to others only for a short period. Actually, the limitations put the private sphere in risk as encrypted information was easy to decrypt by anyone after some years.

Conclusions

Recent development of the Internet and data processing technologies present the risk of a new level of globalized user profiling, as well as new ways of collaborations in order to protect one's privacy. Privacy Enhancing Technologies are developed to change the way how the Internet reveals personal information, but their acceptance is largely different from country to country.

As Regan¹² shows in her analysis, privacy policy harmonization attempts of the last 30 years were moving slowly due to such cultural differences. Without understanding these differences, harmonization efforts of data protection legislation remain cumbersome.

Analysis of these differences is largely needed, however, before starting work on such surveys, let me remind the reader to the effect public opinion surveys could have in political policy making.¹³

References

- Al-Mutadi, Jalal, Karpadia Apu, Mickunas M. Tennis, Naldurg Prasad, *Socializing in the mist: privacy in digital communities*, Paper delivered at the Ubiquitous Computers Conference 2002.
- Bissias, G. D., Liberatore, M., Jensen, D., Levine, B. N., *Privacy Vulnerabilities in Encrypted HTTP Streams*, PET 2005, Cavtat, Croatia, May 30-June 1, 2005.
- Gandy Oscar, H., "Public opinion survey and the formation of privacy policy", *Journal of Social Issues*, vol. 59, n.2, 2003.
- Kostakos Vassillis and Little Linda, "The social implications of emerging technologies: editorial", *Interacting with computers*, vol. 17, 2005.
- Liberatore, M., B. N. Levine, *Inferring the Source of Encrypted HTTP Connections*, CCS2006, October 2006
- Lyon David, "Everyday Surveillance: personal data and social classifications", *Information, communication and society*, vol.5, n.2, 2002.
- Majtenyj Laszlo, "Ensuring data protection in East-eentral Europe", *Social Research*, vol. 69, n.1, 2002.
- Marx, Gary T., "A tack in the shoe: neutralizing and resisting the new surveillance", *Journal of Social Issues*, vol. 59, n.2, 2003.
- Marx, Gary T., "Murky conceptual waters: the public and the private", *Ethics and Information technology*, vol.3, 2001.
- Regan Priscilla M., "Safe harbours or free frontiers? Privacy and Transborder Data Flow", *Journal of Social Issues*, vol. 59, n.2, 2003.
- Turrow Joseph and Hennessy Michael, "Internet privacy and institutional trust: insights from a national survey", *New media and Society*, vol. 9, n. 900, 2007.

12 See Regan Priscilla M, *Safe harbours or free frontiers? Privacy and Transborder Data Flow*, *Journal of Social Issues*, vol. 59, n.2, 2003.

13 See a detailed analysis regarding privacy policy in Gandy Oscar, H., *Public opinion survey and the formation of privacy policy*, *Journal of Social Issues*, vol. 59, n.2, 2003.