Notes for the panel “Privacy Issues in WWW and Data Mining”[*]

Pierangela Samarati[†]
Computer Science Laboratory
SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025, USA
web: http://www.csl.sri.com/∼samarati
email: samarati@csl.sri.com

1 Position by Pierangela Samarati

The privacy problem and the World Wide Web

The increased power and interconnectivity of computer systems available today provide the ability of storing and processing large amounts of data, resulting in networked information accessible from anywhere at any time. It is becoming increasingly easier to collect, exchange, access, process, and link information. In this global picture, people lose control of what information others collect about them, how it is used, and how, and to whom it is disclosed. While before, when releasing some information (be it our health situation to a doctor or our credit card number to a restaurant waiter) we needed to trust a specific person or organization, we now have to worry about putting trust, or some control, over the entire network. It is therefore inevitable that we have an increasing degree of awareness with respect to privacy. Privacy issues have been the subject of public debates and discussions and many controversial proposals for the use of information have been debated openly. In the United States as well as in many European countries, privacy laws and regulations are being demanded, proposed and enforced, some still under study and the subject of debates.

A commonly accepted definition of privacy refers to privacy as the “right of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others.” As we try to spell out the privacy problem with respect to the World Wide Web we can distinguish different aspects, including the classical problem of protecting the confidentiality of information when transmitted over the network (for instance in electronic commerce when we communicate a credit card number over the web); the problem of protecting web surfers from “being observed” as they navigate through the network; the problem of controlling the use and dissemination of information collected or available through the web; and the problem of protecting against inference and linking (computer matching) attacks, which are becoming easier and easier because of the increased information availability and ease of access as well as the increased computational power provided by today’s technology. Although we recognize the importance of providing

[*] 12th IFIP WG11.3 Working Conference, Chalkidiki, Greece, August 1998. The work of Pierangela Samarati was supported in part by the National Science Foundation under grant ECS-94-22688 and by DARPA/Rome Laboratory under contract F30602-96-C-0337.

[†] On leave from Università di Milano.

communication secrecy, we will not discuss this problem any further. We will focus instead on privacy issues concerning information gathering and dissemination.

Privacy issues in data collection and disclosure

Information about us is collected every day, as we join associations or groups, shop for groceries, or execute most of our common daily activities. It has been estimated that in the United States there are currently about five billion privately owned records that describe each citizen's finances, interests, and demographics. Information bureaus such as TRW, Equifax, and Trans Union hold the largest and most detailed databases on American consumers. There are also the databases maintained by governmental and federal organizations, DMVs, HMOs, insurance companies, public offices, commercial organizations, and so on. Typical data contained in these databases may include names, Social Security numbers, birth dates, addresses, telephone numbers, family status, and employment and salary histories. These data often are distributed, or sold. This dissemination of information has been in some cases a matter of controversy (remember the open debates about the plan of America On Line to provide telephone numbers of its subscribers to "partner" telemarketing firms, which resulted in AOL canceling the plan). In other cases, this dissemination of information is becoming common practice. In some states (Texas is an example) it is today possible to get access to both the driver's license and license plate files for a $ 25 fee. Although one may claim that information these databases contain is officially public, the restricted access to it and expensive processing (in both time and resources) of it represented, in the past, a form of protection. This is less true today. Concerns are voiced by individuals who are annoyed by having their phone numbers and addresses distributed, resulting in the reception of junk mail and advertisement phone calls. Even more of a concern is that these data open up the possibility of linking attacks to infer sensitive information from data that are otherwise considered "sanitized" and are disclosed by other sources.

Even if only in statistical, aggregate, or anonymous form, released data too often open up privacy vulnerabilities through data mining techniques and computer matching (record linkage). Tabular and statistical data are vulnerable to inference attacks. By combining information available through different interrelated tabular data (e.g., Bureau of Census, Department of Commerce, Federal and Governmental organizations) and, possibly, publicly available data (e.g., voter registers) the data recipient may infer information on specific microdata that were not intended for disclosure. Anonymous data are microdata (i.e., data actually stored in the database and not an aggregation of them) where the identities of the individuals[1] to whom the data refer have been removed, encrypted, or coded. Identity information removed or encoded to produce anonymous data includes names, telephone numbers, and Social Security numbers. Although apparently anonymous, however, the de-identified data may contain other identifying information that uniquely or almost uniquely distinguishes the individual. Examples of such identifying information, also called *key variables*, or *quasi-identifiers*, may be age, sex, and geographical location. By linking quasi-identifiers to publicly available databases associating them to the individual's identity, the data recipients can determine to which individual each piece of released data belongs, or restrict their uncertainty to a specific subset of individuals [4].

The large amount of information easily accessible today and the increased computational power available to the attackers make inference and linking attacks one of the serious problems that should be addressed. The threat represented by inference and linking attacks is also of great concern because

[1] For simplicity, we refer to the entity to whom information refers as an individual. This entity could, however, be an organization, association, business establishement, and so on.

of the fact that statistical, aggregate, and anonymous data are often exempted from privacy protection regulations. More than others, these data may therefore open up the possibility of potential misuses.

Anonymity issues when surfing the web

While in some cases we know that data about us are collected, but we may not have any control about their use and dissemination; in other cases we may not even be informed that data about us are being collected and distributed. Many people surf the web under the illusion that their actions are private and anonymous. On the contrary, every move they make throughout the net and every access they request are observed and possibly recorded [2]. It is common practice for web servers to maintain a log file recording requests to URLs stored at the server. Each time we hit a web page, the web server records the following information: name and IP address of the computer that made the connection, username (if HTTP authentication was used), date and time of the request, name (URL) and size of the file that was requested and time employed for downloading it, status code or any errors that may have occurred, web browser used, and the previous web page that was downloaded by the web browser (*refer* link). The refer link tells the server the page at which we were looking prior to making the request (i.e., the page "we came from"). One of the reasons for justifying the passing and recording of such information is to allow servers to chart how customers move through a site, and to check the effectiveness of advertisements (as advertisers can control "from where" visitors to their pages arrive). The refer information itself can be seen as a violation of the surfer's privacy, and some more serious concerns arise from information that can be inappropriately leaked through it. Web search engines, such as Lycos, encode the user's search query inside the URL. This information is sent along and stored in the refer link. This means that the server not only knows where we came from, but also what we were looking for. More of a concern is the fact that the URLs fetched from one site using cryptographic protocols (e.g., SSL) may be sent to the next site contacted over an unencrypted link. Thus, for instance, our credit card number that we thought protected because it was communicated over an encrypted link may be communicated unencrypted to other sites. Another threat to surfers' privacy is represented by cookies. A cookie is a block of ASCII text that a web server can pass into a user's instance of a browser and that is then sent to the server (and back again to the browser) along with any subsequent request by the user. Cookies, while providing advantages such as the user's customization, also allow the server to track down a user through multiple access requests to the server and possibly (if cookies are passed among servers) through the entire network. In this sense, cookies represent threats to surfers' privacy.

Data recording information about users' surfing activities over the network are called *navigational* or *transactional* data. Privacy regulations (such as the Electronic Communication Privacy Act) do not generally restrict the use of transactional data; they protect only its content but not its existence. This implies that a service provider can disclose transaction information without the subscriber's consent.

Users concerned with privacy and wishing to anonymously surf the network can today do so by using anonymizing servers. Anonymizing servers act as proxies for the user. Instead of connecting directly to the server they wish to access, users connect to the anonymizing server and pass it the desired URL. The anonymizing server removes a user's identifying information from the request and submits it. The reply also passes to the user through the anonymizing server. In this way the web server of the URL to be accessed receives the request as coming from the anonymizing server. It is worth noticing that in this case the anonymizing server has the ability to observe and record the user's requests. Users need therefore to trust the anonymizer to provide the desired anonymity.

In June 1997, the Electronic Privacy Information Center reviewed 100 of the most frequently visited web sites. The purpose of the review was to examine the collection of personal information and the application of privacy policies by web sites. In December 1997, Bill Helling performed the same survey on the same sites to see whether the situation had changed. Some interesting numbers were reported by EPIC [1] and by Helling [3] as a result of these reviews (numbers reported by the later survey appear in parentheses):

- 49 (57) sites collected personal information (such as name, address, e-mail address) through on line registrations, mailing lists, surveys, user profiles, and so forth. The review could not determine whether the collected information was used for linking data with other databases. Such linking has been found to be performed in some cases (for instance, by America On Line).
- Only 17 (29) sites had explicit privacy policies. Among those, some had policies considered inadequate, some reasonably good. EPIC reports that only a few were easy to find and, although some were considered reasonably good, none of them was considered to meet the basic standards for privacy protection. Helling notes that the sites that later added a privacy policy seemed to make this policy easier for users to locate.
- Only 8 sites provided some ability to the users to limit secondary use of their personal information. This ability is limited to the possibility of specifying whether the collecting organization will be authorized to share (or sell) the information to a third party.
- No site allowed users to review information collected about them. As an exception the Firefly site allowed users to create, access, and update their own personal profile.
- 24 (30) sites enabled cookies. According to [3], 16 of the 30 sites collecting cookies passed the cookie on the home page, before the user could read or link to any explanation. Moreover, at least 7 of the cookies passed on the home page were third-party cookies.

Specifying privacy constraints

Privacy laws and regulations are currently being enforced, and new laws are still under study. They establish privacy policies to be followed that regulate the use and dissemination of private information. A basic requirement of a privacy policy is to establish both the responsibilities of the data holder with respect to data use and dissemination, and the rights of the individual to whom the information refers. In particular, individuals should be able to control further disclosure, view data collected about them and, possibly, make or require corrections. These last two aspects concerning the integrity of the individual's data are very often ignored in practice (as visible from the results of the EPIC survey).

The application of a privacy policy requires corresponding technology to express and enforce the required protection constraints, possibly in the form of rules that establish how, to/by whom, and under which conditions private information can be used or disclosed. With respect to the specification of use and release permissions, authorization models available today prove inadequate with respect to privacy protection and, in particular, to dissemination control or protection by inference. Features that should be provided in an authorization model addressing privacy issues should include

- *Explicit permission.* Private and sensitive data should be protected by default and released only by explicit consent of the information owner (or a party explicitly delegated by the owner to grant release permission).

- *Purpose specific permission.* The permission to release data should relate to the purpose for which data are being used or distributed. The model should prevent information collected for one purpose from being used for other purposes.

- *Dissemination control.* The information owner should be able to control further dissemination and use of the information.

- *Conditional permission.* Access and disclosure permissions should be limited to specific times and conditions.

- *Fine granularity.* The model should allow for permissions referred to fine-grained data. Today's permission forms for authorizing the release of private information are often of a whole/nothing kind, whereby the individual, with a single signature, grants the data holder permission to use or distribute all referred data maintained by the data holder.

- *Linking and inference protection requirements.* The model should allow the specification and enforcement of privacy requirements with respect to inference and linking attacks. Absolute protection from these attacks is often impossible, if not at the price of not disclosing any information at all. For instance, given some released anonymous microdata, the recipients will most certainly always be able, if not to determine exactly the individual to whom some data refer, to reduce their uncertainty about it. Privacy requirements control what can be tolerated, for instance, with respect to the size of the set to whom this uncertainty can be reduced [4].

It is worth noticing that simple concepts, traditionally applied in authorization models, become more complicated in the framework of privacy. An example is the concept of information owner. The answer to this question is not easy and perhaps belongs more properly to the public policy domain. For instance, there have been open debates concerning whether a patient or the hospital owns the information in the patient's medical records maintained by the hospital. Perhaps the notion of owner as traditionally thought does not fit in such context and instead should be revised or substituted by one or more other concepts expressing the different parties involved (data holder vs. individual). A good privacy model should allow the expression of these different parties and of their responsibilities. To the public policy domain will then belong the answer as to how to express such responsibilities (for instance, whether the specification of privacy constraints must remain with the data holder, the individual, or both).

Conclusions

The protection of privacy in today's global infrastructure requires the combined application solution from technology (technical measures), legislation (law and public policy), and organizational and individual policies and practices. Ethics also will play a major role in this context. The privacy problem therefore covers different and various fields and issues on which much is to be said. These notes are far from being complete in that respect. As society discusses privacy concerns and needs, it is clear that research is needed to develop appropriate technology to allow enforcement of the protection requirements.

While stressing the importance of protecting privacy, it is also fair to mention that there are trade-offs to be considered. With respect to anonymity of web surfers, for example, complete and absolute privacy conflicts with the basic requirement of accountability, which demands that users be accountable for actions they execute. Just as we would like not to be consistently observed and recorded while we navigate through the network, it is also true that we would like to be able to determine who accessed our site if, for instance, some violations are being suspected. With respect to data dissemination control and protection from inference and linking attacks, cases may exist where privacy can be (partly) sacrificed in favor of data availability. Let us think for example about data disclosed for scientific research purposes, or about the desire of having globally accessible medical databases so that an individual's medical history be available immediately in case of an emergency, wherever or whenever this might occur. A satisfactory response to these trade-offs may come from the development of new and better technologies. For instance, the application of new measures to protect against inference and linking attacks can allow the satisfaction of data privacy requirements while at the same time maximizing data sharing and availability. Much research needs to be done in this field.

References

[1] Electronic Privacy Information Center. *Surfer Beware: Personal Privacy and the Internet.* http://www.epic.org/reports/surfer-beware.html.

[2] Simson Garfinkel and Gene Spafford. *Web Security & Commerce.* O'Reilly and Associates, Inc., 1997.

[3] Bill Helling. Web-site sensitivity to privacy concerns: Collecting personally identifiable information and passing persistent cookies. *First Monday*, 3(2), February 1998. http://www.firstmonday.dk/issues/issue3_2/helling/.

[4] Pierangela Samarati and Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical Report SRI-CSL-98-03, Computer Science Lab., SRI International, March 1998.