17 SECURITY AND PRIVACY ISSUES FOR THE WORLD WIDE WEB: PANEL DISCUSSION

Bhavani Thuraisingham, Sushil Jajodia,
Pierangela Samarati*, John Dobson, and Martin Olivier

17.1 INTRODUCTION BY BHAVANI THURAISINGHAM

This is the second in a series of panels at the IFIP 11.3 Working Conference on Database and Application Security. While the first panel in 1997 focussed on data warehousing, data mining and security, the panel in 1998 focussed on web security with discussions on data warehousing and data mining.

A data warehouse integrates data from heterogeneous data sources possibly on the web into a single repository so that users can query to make decisions effectively. Data mining is the process of posing queries and extracting information previously unknown possibly form warehouses. The advent of the web together with data warehousing and mining tools is a serious threat to privacy and security of individuals.

The panel positions include discussions of data warehousing and data mining security aspects as well as legal and social aspects of web security. Appropriate privacy laws as well as policies are needed to protect the privacy of individuals. This is a major problem as web id international and different countries have

*Pierangela Samarati is on leave from Università di Milano. 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.

S. Jajodia (ed.), *Database Security XII*

different privacy laws. These heterogeneous policies may have to be resolved possibly to have a uniform policy for the web. There was also interest with the IFIP group to keep in contact with the security group at the World Wide Web Consortium and hopefully to influence the developments with this consortium.

The panelists were: Sushil Jajodia from The MITRE Corporation, Pierangela Samarati from SRI, John Dobson from University of New Castle Upon Tyne and Martin Olivier from Rand Afrikaans University. The positions of the panelists are described below.

17.2 POSITION BY SUSHIL JAJODIA: ACCESS CONTROL IN DATA WAREHOUSES

Generally, the driving force behind the implementation of a data warehouse is the goal of providing a more complete picture of an organization's operations to support management decisions. Although the security concerns for a data warehouse are the s ame as those for any other information systems (integrity, access control, authorization, privacy, and confidentiality), data warehouses present some unique and challenging issues.

Identifying and implementing an access control policy for a data warehouse involves a number of unique challenges. One is the dissonance between access control schemes for data models supported by operational DBMSs and those provided by data warehouse. For example, the relational model is the predominate data model in use today, while decision support systems tend to exploit analytical opportunities offered by non-traditional data models such as the star, temporal, snow flake, or multidimensional data models. The general lack of representation models for defining access controls further frustrates any process for deriving appropriate access controls at the data warehouse level from those used at the operational database level.

In practice, the specification of security policies at the DBMS level is very rudimentary, and organizations rarely document their information system security policies. Finally, users of the operational systems are not the same as the users of the data warehouse, so an access control policy used for an operational system may have little resemblance to one appropriate for the data warehouse level.

Below, we examine issues related to the access control in data warehouses. We have taken a pragmatic approach. By its very nature, a data warehouse creates a conflict between data availability and security [19]. On the one hand, the goal of every data warehouse is to make available to all concerned the information they need, and too much security may have the consequence that users do not have access to all the information that is necessary to do their job. On the other hand, an organization needs to ensure that this same valuable data is not exposed to unauthorized individuals or corrupted by hostile parties. Too little security may mean that users may access information through data warehouse that they cannot access to directly from the sources and, moreover, certain important data may not be made available to the data warehouse by the sources. Therefore, it is important that the correct balance between avail-

ability and security is maintained so that all users that could benefit from some information will have access to it.

Controlling access to a data warehouse is particularly important since the data warehouse encompasses data from many systems and contributes to decision-making across organizational boundaries. In fact, access controls to a data warehouse need to be considered at a number of levels [Ross96]:

- Who will have access to the processes that extract the operational data?
- Who has access to the extracted data and the processes that transform the extracted data into a format suitable for inclusion in the data warehouse?
- Who will access to the data in the data warehouse itself? The ease of access to large amounts of data raises concerns about attaching the appropriate level of security without inhibiting analysis.

In a data warehoue, there is a Warehouse Administrator who should be responsible for deciding which users can execute which processes to extract what operational objects. It should normally be the case that if a user has the privilege to extract data from some operational objects, then that user has the read (or select) privilege to the operational objects in the first place.

Who should control to the extracted data and to the processes that transform the extracted data into a format suitable for inclusion in the data warehouse? Once again, the warehouse administrator should be responsible for designating a small group of privileged users who have access to the extracted data and to the processes that transform the extracted data into a format suitable for inclusion in the data warehouse. The user who creates an object to be stored in the data warehouse becomes the owner of the object and is responsible for deciding which subjects are to have what privileges on the objects.

Who will access to the data in the data warehouse itself? In a relational on-line analytical processing (ROLAP), we see the creation of a star schema as being analogous to defining a view is a relational DBMS. The creator of the star schema becomes of the owner of the star schema and can decide who is to have what types of accesses on his/her object. Just as different views can be defined for different users for security reason, different star schemas can be defined for security as well.

In a multidimensional on-line analytical processing (MOLAP), an entire cube is materialized, and the analysis tools are applied directly on the materialized cube. We view each materialized cube as an authorization object. The user creating the data cube must have the read privilege on the underlying detail data, and gets to decide who has subsequent access to the cube. No attempt is made to hide parts of a cube from the analyst.

17.3 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 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[2] 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 [25].

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 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.

[2]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.

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 [11]. 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 [10] and by Helling [14] 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 [14], 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 [25].

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.

17.4 POSITION BY JOHN DOBSON: WHY IS INFORMATION PRIVACY SO HARD?

The best definition of privacy that I know is one due to Joachim Biskup that defines it in terms of role separation: an individual in society takes on a number of roles, and privacy is the right to expect society to respect the individual's chosen separation between these roles.

One problem is that social roles are constantly renegotiated in conversations and changing relationships, or are subverted by legislation; therefore any view of privacy that assumes roles are fixed or immutable is likely to be inadequate. Unfortunately, many computer systems take this view.

A second reason is that the location of the public/private boundary is culturally determined, and therefore a system developed in one culture with one set of assumptions about the location of that boundary is likely to prove unfit for purpose in another culture with another set of assumptions.

A third reason is that we don't know yet how to transfer our understanding of normal social relationships to computer-mediated social relationships. Every time you engage in a social relationship you take a privacy risk. In circumstances we understand, we can evaluate this risk (at least subjectively) in making the decision how much to commit to the relationship. In computer-mediated relationships we don't yet have enough experience to know how to do this. Furthermore, part of this risk evaluation depends on the existence in the social world of recourse: we can employ the courts and other social sanctions and actions if we feel we have been betrayed. Again, in computer-mediated relationships we don't know how to do this.

More generally, there is a distinction between security and privacy which throws an interesting light on the issue, and it is this: in security, role definitions have been institutionalised whereas in privacy, role definitions have not been (and probably cannot be) institutionalised. By this I mean that our understanding of security depends on our knowing who is supposed to know or have access to what, this knowledge being based on public knowledge of assignment of role; whereas our understanding of privacy depends on the individual's choice of which roles to assume, this being an individual matter and not necessarily open to public knowledge. The importance of the kind of institutionalisation required for security is that mechanisms to support it can be made part of the social or technical infrastructure underlying the institution, whereas this is just not possible for something like privacy that almost by its very definition has to remain structural. In fact any information system has the effect of forcing the institutionalisation of role, and that is why they can be so dangerous to privacy.

A related way of putting this is that in terms of a formal logic, infrastructural definitions (e.g. of security) can be expressed in extension and can therefore be the subject of mechanical interpretation whereas structural definitions (e.g.

for privacy) can only be expressed intensionally, and must therefore remain in the domain of policy and legislation.

17.5 POSITION BY MARTIN OLIVIER: PERSONAL PRIVACY

A human being's personal privacy refers to his/her ability to limit collection and use of his/her personal information. Where such limitations cannot be determined by the individual, suitable mechanisms have to exist to ensure that collection and use will be adequately limited. While a significant amount of work has been done about what constitutes *adequate limitations* (see below), hardly any work has been done on suitable (technical) mechanisms to provide any guarantees about such limitations.

It is the contention of this author that such mechanisms are not only required, but that it is also technically feasible to develop them.

General privacy principles

For computing purposes personal privacy may be split into communications privacy and database privacy. *Communications privacy* has to ensure the privacy of personal information during communication. Privacy mechanisms may enable the individual to control what personal information is communicated, what information is collected by the party it communicates with and for what purpose such information is collected. In addition to these controls, communication privacy requires communication security mechanisms to ensure integrity of communicated information and confidentiality of such information against third parties.

Database privacy has to ensure proper use of personal information once it has been collected. Fundamental to such proper use is use of the collected information only for the purpose for which it has been collected. Three principles govern proper use of personal information: The *need to know privacy principle* is similar to the need to know security principle, but restricts a subject to access an individual's personal information to *when* the subject needs the access to *that* individual's information. The *acceptable use* privacy principle mandates that no-one should be able to use information for purposes other than for what it had been collected. In particular does this principle prohibit comparison or aggregation of or derivation of information from personal data collected for incompatible purposes. *Integrity of personal information* has to ensure that information is correct, timely and up to date.

Privacy protection is both a personal and fundamental right of all individuals. Individuals have a right to expect that organisations will keep personal information confidential. One way to ensure this is to require that organisations will collect, maintain, use, and disseminate identifiable personal information and data only as necessary to carry out their functions.

Privacy in practice

The principles given above form the basis of ethics viewpoints on privacy [22, 2, 6, 18, 20, 26]. Laws are also based on these principles. The US Privacy Act of 1974, for example, limits government (federal) use of personal data to relevant and necessary data to accomplish the purpose of the concerned federal agency [22].

Processing of personal information is allowed according to the European Data Protection Directive if the concerned subject has unambiguously consented to the processing for which notification has been given as to the purpose for which the information is sought [20]. Otherwise "the processing of personal data must ... be necessary with a view to the conclusion or performance to a contract binding on the data subject, or be required by law, by the performance of a task in the public interest or in the exercise of official authority, or by the interest of a natural or legal person provided that the interests or the rights and freedoms of the data subject are not overriding" [8]. The directive requires "appropriate safeguards" to ensure that personal information is not compromised, but does not specify the nature of such safeguards.

Correctness, and legitimate use of personal information is addressed by the right of any individual of "... access to data relating to him which are being processed, in order to verify the accuracy of the data and the lawfulness of the processing" [8].

Note that, once information has been collected for "specified, explicit and legitimate purposes" it may not be " further processed in a way incompatible with those purposes" [8, Article 6.1(b)].

As a national example, the Dutch Data Registration Act [27] also limits use of personal information to the purpose for which it has been collected (Article 6), information has to be obtained lawfully (Article 5) and individuals normally should be able to access their personal information (Article 29). Not only are individuals able to request the contents of such records, but also the origin. The law also requires the technical and organisational protection of such information against unauthorised access or modification. The law does, however, primarily expect the individual to monitor application of the law [26].

Some laws even go so far as to prohibit the collection of some personal information. The European Data Privacy Directive requires that "Member States shall prohibit the processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, and the processing of data concerning health or sex life" [8, Article 8.1] except in a small number of enumerated cases.

Laws also realise the power of computers — in particular the power to derive information that would not have been possible without their ability to aggregate and compare large amounts of data. The US Computer Matching and Privacy Protection Act of 1988 states that "agencies must follow specific procedures when engaging in the automated comparison of Privacy Act databases on the basis of certain data elements" [22].

Automated privacy mechanisms

Very little has been done to use technology to enhance privacy. Almost all such attempts focus on the individual's ability to limit the collection of private information. It has been argued that technology should be used to allow the individual to specify personal preferences as to what information may be collected [15, 5]. How a specific individual's personal information will be handled, will be negotiated before any interaction between an individual and an organisation occurs. If such contact is on-line (such on the World Wide Web), negotiation may even be automated. How can an organisation be trusted that it will honour its privacy undertakings? Here it has been suggested that organisations may be 'certified' by some mutually trusted body [9, 3]. Various possibilities then exist if the organisation does not adhere to its undertaking.

Technical work has also been done that enables an individual to withhold information from an organisation by interacting anonymously with the organisation. This ranges from 'anonymysers' on the World Wide Web, anonymous remailers on the Internet, electronic forms of cash and other secure payment protocols.

There are, however, many instances where collection of private information cannot be avoided. And, once personal information has been collected, it needs to be properly protected.

Traditional security mechanisms form the first defence against privacy violations. But traditional security is not enough to ensure privacy. The majority of the privacy violations mentioned in [21, 17] were committed by authorised users of the system. Anderson [2] rightly points out that the privacy problem is compounded if the value of information increases or the number of people that have access to it increases — the former because the incentives to misuse information are that much higher; the latter because the potential to find an unscrupulous individual is simply higher in a larger group of (often remote) users. Aggregating personal information in a centralised database increases both the value of the information and, usually, the number of people who have access to it. Someone who is allowed access to some individual's personal information should not necessarily be allowed access to the same information of other individuals.

Solutions

Data privacy needs mechanisms that will limit use of personal information according to some privacy policy. Many privacy policies already exist: a quick survey of many corporate (and other) Web-sites will show that many of them have a privacy policy in place. These policies often assume that individual employees of such organisations are trustworthy and will adhere to the accepted policy. The examples cited in [21, 17] show that this is not, in general, true for *all* employees. To address this problem, policies aimed at limiting the damage an authorised individual may do need to be considered. Note that perfect policies do not exist: consider the doctor who obtains personal information

shares this information with an individual not concerned with the case. This violates privacy, but neither the collection of the personal information, nor the sharing of the information can be controlled by technical means. We therefore need policies that technically limit the possibility of violations and then use professional, ethical and legal means to address the remaining problem.

The starting point is obviously traditional security mechanisms: by limiting access to individuals or roles who need to know the information, already avoids many potential violations.

Various possibilities limiting users' abilities to browse personal information also exist: a tax inspector needs access to personal tax information; however, no need exists for such an inspector to be able to access all individuals' tax records. This seems to require partitioning mechanisms based on users: with any particular user (or subject) only having access to a subset of individuals' information, the potential to violate privacy is again reduced. Determining suitable subsets depends on the application: A doctor may, for example, only access medical information of patients treated by him/her. The subset of taxpayers whose information a tax inpspector may see, may conceptually be determined randomly.

Possible 'technical' privacy policies to limit privacy violations further, may focus on the partitioning of information in databases based on the purpose for which such information has been collected. Limiting an individual's access to a single partition (or a few partitions) may be the first step to limiting users' abilities to acquire unneeded information. In particular does limiting computerised correlation and aggregation of information across partitions have potential.

Once such technical policies have been devised, it becomes possible to design mechanisms to support these policies.

Ensuring privacy of information usually does not hold immediate benefits for the concerned organisation, as secrecy of organisational confidential information does. The reasons why organisations may consider implementing privacy controls are twofold: Firstly, they may be forced by law to ensure privacy. Secondly, having such controls in place may be indirectly beneficial to the organisation, in the same way that steps taken to prevent pollution may be. In both cases it is necessary to audit the controls to ensure that they are adequate and effective.

References

[1] An introduction to multidimensional database technology. KenanSystems Corporation, `http://www.kenan.com/acumate/mddb_toc.htm`.

[2] RJ Anderson. Patient Confidentiality — At Risk from NHS Wide Networking. Health Care '96.

[3] A Blackburn, L Fena and G Wang. A Description of the eTRUST Model. in Chapter 5, [6].

[4] Surajit Chaudhuri and Umeshwar Dayal. An overview of data warehousing and OLAP Technology, ACM SIGMOD Record, Vol. 26, No. 1, March 1997, pages 65-74.

[5] LF Cranor. The Role of Technology in Self-regulatory Privacy Regimes. in Chapter 5, [6].

[6] WM Daley and L Irving. *Privacy and Self-Regulation in the Information Age*, US Department of Commerce, Washington, DC, June 1997.

[7] Barry Devlin. Data Warehouse from Architecture to Implementation. Addison-Wesley, 1997.

[8] *Directive 95/46/EC on the Protection of Individuals With Regard to the Processing of Personal Data and on the Free Movement of such Data*, 24 October 1995, European Union.

[9] E Dyson. Labeling Practices for Privacy Protection. In Chapter 5, [6].

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

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

[12] Jim Gray, Adam Bosworth, Andrew Layman, and Hamid Pirahesh. Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals. Technical Report MSR-TR-95-22, Microsoft Research, Redmond, WA, November 1995.

[13] Venky Harinarayan, Anand Rajaraman, and Jeffrey D. Ullman. Implementing data cubes efficiently. Proc. ACM SIGMOD International Conf. on Management of Data, 1996, pages 205-216.

[14] 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/`.

[15] LJ Hoffman and KA Metivier Carreiro. Computer Technology to Balance Accountability and Anonymity in Self-regulatory Privacy Regimes. In Chapter 5, [6].

[16] W. H. Inmon, J. D. Welch, and Katherine L. Glassey. Managing the Data Warehouse. John Wiley & Sons, Inc., New York, 1997.

[17] *IRS Systems Security: Tax Processing Operations and Data Still at Risk Due to Serious Weaknesses*, United States General Accounting Office, Washington, Document GAO/AIMD-97-49, 1997.

[18] S Jajodia. Managing Security and Privacy of Information. *ACM Computing Surveys*, **28**(4es), 1996.

[19] Ralph Kimball. Hackers, Crackers, and Spooks, Ensuring that Your Data Warehouse is Secure. DBMS, April 1997, pp. 14-16.

[20] I Lloyd. An outline of the European Data Protection Directive. *The Journal of Information Law and Technology*, 31 January 1996, `http://elj.warwick.ac.uk/elj/jilt/dp/intros/`.

[21] *National Crime Information Center: Legislation Needed to Deter Misuse of Criminal Justice Information*, United Stated General Accounting Office, Washington, Document GAO/T-GGD-93-41, 1993.

[22] *Options for Promoting Privacy on the National Information Infrastructure*, Information Policy Committee, Information Infrastructure Task Force, Washington, 1997.

[23] Arnon Rosenthal, Paul A. Dell, Pamela D. Campbell. Integrity and Security in Data Warehousing. AFCEA, 1997.

[24] Steven J. Ross. Control Issues in Data Warehousing. Infosecurity News, July/August 1996, pp. 22-24.

[25] 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.

[26] JHJ Terstegge. Personeelsinformatiesystemen en Privacybescherming. In F de Graaf e.a. (Red), *Handboek Privacybescherming Persoonsregistratie*, Paragraaf 2107 HD Samson, Tjeenk Willink, Alphen aan de Rijn, 1982 (Supplement 19, September 1995).

[27] *Wet Persoonsregistrasies*, Nederland, 1988.