COMPUTERS AND PRIVACY

by

Richard L. Garwin

IBM Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

(914) 945-2555

(also
Andrew D. White Professor-at-Large,
Cornell University;
and
Adjunct Research Fellow,
Harvard University;
and
Adjunct Professor of Physics,
Columbia University)

Presented at the Embassy
of the
United States of America
London, ENGLAND

January 18, 1983

It pleases me greatly to continue the tradition of speaking to an audience in London at the invitation of the U.S. Embassy. I am grateful also for the sponsorship of the Science Policy Foundation. I fear, however, that I shall disappoint any of you who expect an analysis of the Data Protection Act, laid before Parliament last month. My talk, rather, stems from a long-standing interest in a topic on which I wrote under the title, "Some Aspects of Technology and Privacy," published as a staff report of the Senate Select Committee on Intelligence, Volume IV, pages 109-119 (April 23, 1976) (as "Intelligence and Technology"). In that paper, I sketch the evolution of technological threats and aids to privacy in regard to the Fourth Amendment (U.S. Constitution) Right to be secure in one's person, papers, and home, specifically in 3 fields: hidden microphones and cameras, intercept of voice and non-voice electrical communications, and the creation and search of large files. The Fourth Amendment (of the 10 constituting the Bill of Rights, in force since 1791) reads:

> "The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no warrants shall issue, but upon probable cause, supported by Oath or affirmation, and particularly describing the place to be searched and the persons or things to be seized."

This early and fundamental legal support of the right to privacy has been supplemented by specific legislation, provoked by experience in some cases, but in others responding to the potential of new technology. Thus the Federal Communications Act of the 1930s forbids unauthorized wiretapping of telephone conversations, and that protection is extended (where U.S. law runs) to the acquisition of inter-personal communications from which intelligence can be "aurally acquired," even from a radio or microwave relay link in a telephone system. However, the transmission of data (facsimile, teletype, inter-computer communication) is not so protected.

In the present talk, I shall limit myself to computers and their potential to protect or impair privacy. Because basic law and custom are so different in the United States and the United Kingdom, I cannot usefully comment specifically on the Data Protection Act, but I can provide some background from the U.S. scene.

As in my earlier papers, I emphasize that I give you my own thoughts and not an IBM position. In decades of involvement in matters of public policy including national security, health care, environment, and regulation, I have tried as a scientist to provide the tools and information necessary to balance benefits against costs, including the extrapolation to the future of potential costs and benefits.

The desire for privacy varies among cultures. A correspondent to a New York newspaper living in Tokyo was shocked when he received his bank statement on a postcard. The technology of envelopes is well-known to the Japanese; the need for privacy was not. In our society it is no longer regarded as prudent to publish in the local newspaper the fact that the family will be away for a 2-week vacation, since burglars know how to read and have an interest in exercising this skill. Universal access to hotel registers or airplane passenger lists would present this same problem, and others as well.

The laws of a democracy prescribe the means for changing the government; explicitly or implicitly they also limit the means by which a government can stay in power. A government file of personal information which would be regarded as essential in an oligarchy could be anathema in a democracy. The historical U.S. emphasis regarding privacy has been for protection against actions by government, hence the principal focus of these remarks.

In some cases new technology (in the present instance, computers) may aid in preserving privacy against invasion by people or tools. An old example is the use of locks on doors; a newer one is the use of encryption for written communications and for the privacy of information in files. On the other hand, it would be inappropriate to require the individual to go to great cost to preserve his rights if such protection could be obtained at lesser social cost, as by restrictions on the actions of individuals who would intentionally violate these freedoms or whose activities might inadvertently imperil these rights. Thus, the expectation of privacy for the contents of a postcard sent through the mails is quite different from that of a letter in a sealed envelope, and the cost of an envelope is not regarded as an excessive charge for the guarantee of privacy. As the human senses of capabilities of vision, hearing, and memory are expanded by the use of new tools, what is the place for the analogue of better envelopes? The envelope provides legal protection, not durable physical protection-- a distinction to be kept in mind as we distinguish regulation from encryption, for instance.

Blackmail, burglary, unfair personal competition may result from personal information in the possession of an individual willing to use it for criminal ends or even to injure another or to benefit himself at the expense of another in ways not explicitly forbidden by statute. In recent days, an ex-employee of the Federal Reserve Board was found attempting to gain access (presumably for private gain) to the Board's official projections of money supply-- not by wire tapping but by unauthorized remote access to the Board's computer. Such hazards may also result from manual information files and from manual processing, but in reducing cost the computer may change the criminal's cost-benefit analysis. I do not regard such potential ills as fantasy, and it is the responsibility of legislators, governments, industry groups, and individuals interested in the public welfare to make available remedies even before substantial harm is demonstrated-- with care that such remedies not deny society important benefits of new technology.

In the United States, the Freedom of Information Act is an attempt to make available on request any material held by the Federal government which is not exempt from release, while the Privacy Act of 1974 allows the individual in many cases to refuse to provide information, to know of the existence and content of a government file concerning that individual, and to challenge information which may be found in that file. These laws apply to both manual files and those which are computer served. As regards private files, these protections are provided thus far primarily in the Fair Credit Reporting Act and in the Family Educational Rights and Privacy Act.

Let us recall how computers are used for the storage and manipulation of information.

File Technology.

Some examples of current status. Among the early large computerized file-oriented systems are the seat reservations systems now in use by all airlines. The overall system accommodates thousands of flights per day, with a hundred or more seats per aircraft, and can handle reservations months ahead. A reservation can be made, queried or cancelled within seconds from any of thousands of terminals. Some of the records may contain little more than the name of the passenger; others may include a complex continuing itinerary, with hotels, car rental, telephone numbers, and the like.

Several government echelons have files (data bases) for tax purposes. At the city or county level, such data bases may include details about every dwelling in the city; they can be particularly useful in case an overall reassessment is desired, but they may also be used to aid firefighting or police work. Who would argue that public authorities be prevented from obtaining names and telephone numbers of residents of a flat and its neighbors, if a gunman is shooting from the window?

The New York Times Information Bank ("NYTIB") provides at the New York Times building both abstracts and full texts of articles published in that newspaper. From remote terminals, subscribers can search the compendium of abstracts for all articles which have been published in the New York Times and may request copies of the full articles whose abstracts satisfy the search criteria. The abstract searching can be full-text search, i.e. a search on the name "Harold Ickes" might result in a sheaf of abstracts, accompanying stories most of whose headlines say nothing about Ickes, but may refer to Roosevelt.

Full-text search capability is now commonly used in connection with law and legal decisions. In addition to struggling with the often inadequate index to such a corpus, an attorney can undertake a full-text search for statutes or cases which have some characteristics in common with his current concern.

Business and government maintain personnel files-- how else would employees be paid? Individuals have medical records with their physicians, clinics, hospitals, insurance providers, peer review groups, and the like.

All these are file-oriented systems, some of which may retrieve files according to the index system under which they were prepared; others, as we have seen, have a full-text search capability, such that a file can be retrieved in accordance with its content rather than heading. In some cases the file is available for search with any editor; in others, a "transaction manager" program provides limited response to validated requests.

Computer file systems are now in common use for text preparation and editing. A draft letter, report or publication is typed at a terminal connected with a computer or at a stand-alone system. At any time, portions of the draft can be displayed, typed out locally or on a fast printer. The typist can enter corrections into the computer system (including global changes, e.g. to change the group of characters "seperate" every place it may occur into the group "separate", or "Paris" into "Brussels"), can rearrange paragraphs, append additional files, and the like.

The use of computers in all these file applications-- commercial, education, and intelligence-- is motivated by a drive for efficiency, reliability and the capability to retrieve materials at places, times, and by persons other than those who have filed them. Computers at present are not normally used to store pictures or things, but indexes to such collections can as readily be placed in the computer as can any other kind of information. In contrast with a single physical file of paper documents, the computer store never suffers from the document's unavailability because it is on somebody's desk. Multiple copies of a micro-image store can also satisfy the requirement for multiple simultaneous use, but cannot be updated or searched so readily as can a computer store.

Current file technology-- performance and cost. Obviously, concern regarding files and privacy is with the chain of information from collection through storage and retrieval. One worry is that some private or government organization by the expenditure of enough money, could have the capability to "Know everything about everyone" at any time. Because there is no general public right of access to the files of police or intelligence agencies, it is of interest to know what these capabilities might amount to, as a guide to the introduction of safeguards. This week in the U.S., the Los Angeles police department is accused of not having destroyed various intelligence files as promised, so the question is more than hypothetical.

In order to provide some intuitive feeling for the magnitudes involved, consider the storage of full-page, double-spaced text. Such a page may have 30 lines of 65 letters or digits, or about 2000 characters per page. Except as noted, it is assumed that a character requires one "byte" (8 bits) of storage, although by appropriate coding of text, one can store as many as 3 characters per byte.

Using a disk-pack magnetic storage device such as has been commercially available for 6 years or so (typified by the current IBM Model 3380) storage of 2500 million bytes can be obtained for a rental of about $3000 per month or some $1.2 per month per million characters. Such a device can transfer about 3 million characters per second, so it would require 800 seconds to search its entire contents if the logical search device operated at the storage data rate. Search is normally done by a query, looking for an exact match in the data stream as it is brought from the store. Examples of simple queries are: "theft of service" in the case of the legal corpus: "Thatcher/Falklands" in the case of the NYTIB (where the "/" simply means that both "Thatcher" and "Falklands" should be in the same document); "seperate" in the case of ordinary text processing where the properly spelled word "separate" is to be substituted. Such queries against a small data base are handled well by a general purpose computer. Large data bases often have some structure which can be used to reduce by large factors the amount of data which actually has to be searched. Even if the data base has little structure, one could imagine streaming the entire data base past some modest special-purpose electronic device (a "match register") which may detect a match against the query and divert the matching document into a separate store, where it may be brought to the attention of the user. In large production, such a match-register might be bought for $5 in modern integrated-circuit technology. In any case, the cost of special-purpose match-registers would be small compared with the cost of the massive store.

By such techniques, as many queries as are desired may be entered from terminals and simultaneously matched against the entire data stream. If the data base is entirely in this type of storage (at a present cost of $1.2 per month per megabyte, or 12 cents per month per nominal files of 50 typed pages) any query can be answered within 15 minutes. Of course, a single query might lead to many other sequential queries before all the desired facts are at hand, but the time is measured in minutes, not months.

Given that most queries need not be answered in minutes, one can ask the cost of a slower (hence cheaper) system. A commercially available tape library product (the IBM Model 3850-A04) can store 236,000 million characters at a cost of about $28,000 per month (so 9 cents per month per million characters stored). Assuming that this particular device can deliver data at a rate of 3 million characters per second, it would require some 20 hours for such a store to be searched entirely for as many queries as have been presented. The range of cost associated with such a system with current technology and twelve-hour response time thus goes from $2 million per month for a system capable of storing 50 pages on each of 200 million individuals (without encoding) to about $40,000 per month for a system storing the same information on each of 10 million individuals, with the characters compacted into more efficient form for storage.

So much for current technology, in operation all over the world for normal commercial purposes. It serves highly important functions in allowing any organization-- commerce, industry, government, and the professions-- to manage information quickly and accurately. Yet it may have undesirable side effects.

Still in our memory is the use by the Nixon White House of illegal means (including burglary, but not computers) to provide a "psychological profile" on Daniel Ellsberg. An ordinary file drawer would be adequate if one knew long in advance that information would be requested on this particular person. Given the unusual nature of the case and the non-existence of that particular file drawer, it would be technically possible to search all government files for documents which mentioned the name in question. This would bring to light, of course, income tax returns, military service history, all employees for whom social security tax had been paid in the past by the individual in question, names of relatives, and so on. This material could be found if the queries were made available to cooperating individuals with access to files in agencies like the Internal Revenue Service, the Social Security Agency, Selective Service and the like. Additional important information might be available by use of the NYTIB as a commercial subscriber.

Thus the problem in regard to those government agencies with large files of raw data is to ensure that these files are used only in support of the authorized mission of the agency and are not exploited for purposes of improving prospects of incumbent officials in an election, of punishing those on an "enemies list," and the like. It is no longer enough to proscribe the creation of specific dossiers on individuals; it is now possible to recreate such a file from the central file in less than a day, or to answer questions from the central file without ever having a manila folder or file drawer labelled "John Smith." There must therefore be control over the nature of queries asked of the file, of whom, and by whom. It is just as important to ensure that information given freely by individuals is not exploited for unauthorized purposes and is not accessible to unauthorized individuals.

The computer technology which makes possible rapid access to large masses of information also allows in principle for control of access to that information. Measures for preventing illegitimate use of government files must be proposed by the Executive, with advice from equipment manufacturers, organizations experienced in computer use and analysis, from the scientific societies, and from consumer organizations. Such measures could be embodied in Executive Orders. Their adequacy and the need for legislation providing criminal and civil penalties should be the subject of further Congressional hearings and research.

Legal and Procedural Safeguards which are being considered and partially implemented in personal data files are the following:

1. There should be a limitation as to who can keep files on individuals (But clearly the New York Times is allowed to put their own newspaper into computer-readable form. And is it a file on an individual if the individual's name is only mentioned in a larger document?),
2. Individuals should be allowed access to their files (for repayment of the actual cost of search) and to receive the information in the file on them. (But if the file is very large, such access might be made very expensive. On the other hand, if the access were treated like an ordinary query in the example above, the cost might be quite reasonable),
3. The individual should be allowed to write into the file in order to contest the facts or in order to present his own point of view,

4. There must be limitations on those who can gain access to the file or who can receive information from the file.
5. Duplication of the file must be limited and unauthorized access prevented,
6. There should be an indelible record of who has queried the file and what questions were asked, so that failure of access limitations will not go undetected.

Among the safeguards for any system containing large amounts of sensitive data should be adequate requirements for identification of terminals from which queries are being made, identification and authorization of the individuals who query, a complete record of the queries (with terminal and individual identification), adequate security against transmitting large amounts of information and the like. The moment-by-moment execution of these controls on access is the task of the set of computer instructions known as the "operating system."

The design of an adequate operating system is a difficult task, and even the detailed specification of access controls is not simple and must be done with some understanding of what is technically feasible at present. Fundamental to the continued effectiveness of such safeguards is the maintenance of the integrity of the main program which controls the computer. Clearly, the introduction of access controls should not wait for the perfect operating system.

The most secure computer is the one which cannot be used, but the demand for ultimate computer security would take us back to paper files with their own exposures of security and privacy. Proper attention to security and privacy can increase the difficulty and cost of unauthorized access so that it is easier to obtain information in some other way. Indeed, ease of use does not conflict with high security, and difficulty of access in itself does not guarantee privacy.

No matter what the safeguards, individuals might be able to gain access to some information for which they are not authorized. Adequate legislation, criminal penalties, and the enforcement of these laws should deter many who might otherwise try. Data security measures, such as encryption of the file itself, can help also. Indeed, open analysis by all those concerned should lead to an understanding of the protection which may be provided.

What must be particularly guarded against is the misuse of information freely given or collected for authorized purposes and which is then turned to an improper use. Times change, and with them standards. When the U.S. Social Security system was created in the 1930s, it was promised that the Social Security Identification Number (SSN) would be used only for Social Security purposes; it is now demanded by Internal Revenue Service and by many other governmental and non-government organizations. In the 1940s my uncle was pleased that he had improved the efficiency of the large public utility for which he worked (and also helped the employees) by using the SSN as the employee identification number; by 1975 IBM replaced the SSN by an Employee Identification Number in all files and on all materials except those where the SSN is legally required.

Some personal experience

In 1983, the number of personal computers sold each year approximates 1 million and is growing rapidly. It seems reasonable to assume that many of the owners of personal computers want to improve the quality and reduce the drudgery of personal correspondence and record keeping. The first thing they will do is to consolidate their manual address files, so that the names, addresses, and telephone numbers of correspondents may be readily and efficiently obtained. It is far easier to keep such a computer file up-to-date than to change fragmentary manual files. Some of those in this room are in my own "NAMES" file system for these very reasons. And with increasingly fragile human memory, I sometimes use the file to store the name of a correspondent's spouse.

In writing someone one would like to refer to the previous correspondence. In particular, I want to know which of my publications or informal writings I have already sent a correspondent, and I have a MAILOG system which provides this information expeditiously, containing a single line of description for each incoming or outgoing letter.

For electronic mail with those individuals whose computers talk to my computer, I maintain OLDMAIL files, one for each correspondent, but these OLDMAIL files also contain the names of people to whom copies of incoming correspondence were sent. In some cases, these names are not real names but nicknames, or computer user identification, so I have no idea who the people actually are. Within a few years, it will be common for individuals with personal computers to manage their personal correspondence in such fashion. They will also find it useful when visiting London (or New York) to obtain a listing of those of their correspondents living in London, as I have done myself in connection with this visit. What regulatory controls are appropriate to a file used by an individual, or by an individual and staff support for correspondence?

The rights of the individual may be protected by a regulatory approach or by resort to recourse. I tend to favor legislation defining acceptable and unacceptable management and uses of information, and the circumstances and penalties governing recourse. Under this approach, those who maintain or process files (either manually or automatically) would have a statutory duty to observe the regulations, as a defense against recourse. The vast amount of comment already existing on these matters inhibits my adding to this corpus, but I shall close with a few remarks about the virtues of computers in supporting the right to privacy.

I have already observed that a file in the computer is never unavailable because it is "on someone's desk." The solution to that problem with manual files is the replication of the file, which makes it very difficult instantly to deny access to the file, to update or to correct the file, or to purge information from the file-- all facilities agreed by everyone to be fundamental to file management.

In contrast, the problem of computer data protection is limited to the single computer data base, and the techniques of controlled access, passwords, and the like are applicable. However, the information stored on magnetic tape or on a magnetic disk would be available if that storage system were mounted on another computer, except for the fact that the file can be encrypted.

Non-procedural protection--Encryption. From history and fiction, you are all familiar with cipher systems, which by now have attained such a state of security that a nation can broadcast freely day after day most secret information which could be deciphered into plaintext by anyone possessing the key. A cipher system is suitable for such purposes only when the work (manual or computer-aided) required to decrypt without the key is comparable with that involved in "trying all possible keys." Thus for a 7-byte (56-bit) key, there are two to the 56'th power possible keys, (assuming all 256 possibilities for each byte are utilized), and so six times ten to the 16'th power trials at deciphering will have to be conducted to be sure of success. In the 1970s, the U.S. National Bureau of Standards (NBS) invited proposals for a cipher system to be used for protection of non-military data in the U.S.. A modification of a system proposed by IBM was adopted-- the Data Encryption Standard (DES), now provided by several suppliers in the form of a computer program (software) and a chip set (hardware) for various applications including automated banking tellers, file protection, and the like. DES normally uses a 64-bit key of which 56 bits may be independently chosen by the user-- a choice among the 64,000 trillion keys. Thus, if a new key could be tried and tested every microsecond, it would take 3000 years to decipher communications by exhaustive search. Naturally, the user will change key at suitable intervals.

On the other hand, even a perfect system of encryption can be misused. For instance, if one uses as a key 7 consecutive characters beginning at an arbitrary word in a 100,000-word novel, there are only 100,000 keys to be tested, and a system far less capable than one testing a key per microsecond will render the file readable in a short time. ((Sound professional advice on choice and management of keys is desirable; for instance, DES users have been advised not to use a few "weak" keys. However, if keys are selected at random (as by the use of DES itself as a pseudo-random-number generator), there is an insignificant chance of choosing one of a few weak keys. A user skeptical of the randomness of DES should feel perfectly free to add a sub-key of the user's choice to the candidate "random key".))

With proper use of DES, then, the data in the computer can remain encrypted. Remote terminals may communicate securely, with a DES chip enciphering the data before it goes over the telephone line, and deciphering again at the central computer. The information found from the deciphered request processed against the deciphered file is then re-encrypted with the transmission key and sent back to the terminal where it is exhibited as plaintext. It is possible for the "session manager" program to validate the nature of the request, the person requesting, and provide access to the file key. Thus the file key itself need not be compromised by being distributed to large number of individuals.

Another application for encryption is evident to nullify the value of data banks if captured or stolen; it would have been useful for the U.S. Embassy in Iran to have had only encrypted information rather than materials stored in safes.

Data encryption should be more widely used for the protection of valuable or sensitive data.

Summary

Protection of privacy and other human rights is a continuing task. The process should be a continuing one as well. It includes the recognition of problems, the identification of opportunities, the drafting of proposed regulations and legislation, the evaluation of the impact of such rules on interested individuals, and the iteration of this process. Technology can ease these problems, as by the use of data encryption. Not only must there be representation for the individual citizen as data subject, there should be representation for those who will be using new technologies which have not yet found their way into widespread application. Attention must be paid to the non-technical portions of legislation-- clear statements of condition for the issuing of warrants; assignment of individual personal responsibility to government officials and others who violate the law, with substantial criminal and perhaps civil penalties prescribed. Such legislation should not be so narrowly drawn as to fail of protection against present technology or that which may arise in the future. Even appropriately broad legislation will have to be reviewed periodically in the light of experience and new technology to ensure that these rights are protected against the actions of government and of private individuals.