FOUNDATIONS OF COMPUTER SCIENCE
Series Editor: Raymond E. Miller

Principles of Data Security

Ernst L. Leiss

Principles of Data Security

Ernst L. Leiss

University of Houston
Houston, Texas

PLENUM PRESS • NEW YORK AND LONDON

Library of Congress Cataloging in Publication Data

Leiss, Ernst L., 1952-
Principles of data security.

(Foundations of computer science)
Bibliography: p.
Includes index.
1. Computers—Access control. 2. Data protection. 3. Electronic data processing departments—Security measures. I. Title. II. Series.
QA76.9.A25L48 1982 001.64'4 82-22272
ISBN 0-306-41098-2

Preface

With the greatly increased use of electronic media for processing and storing data, data security became an important practical issue. This is especially true for the extensively shared and distributed systems which are more and more being accepted in commercial operations. Essentially, the problem is that of protecting data, including all the implications which this has to the end user as well as the systems or database designer. However, usually the term *data security* refers to protection by technical, i.e., computer science specific, means; if one wants to include issues such as physical security, how to select the group of people who should have authority to perform certain operations, etc., the term *computer security* is more appropriate.

The object of this book is to provide technical solutions to (facets of) the problem of achieving data security. The reader who hopes to find clever recipes which allow circumventing protection mechanisms will, however, be sadly disappointed. In fact, we deliberately kept the presentation of the material at a fairly general level. While the short-term benefit of such an approach may be somewhat smaller, we feel that without a thorough understanding of the fundamental issues and problems in data security there will never be secure systems. True, people probably always considered certain security aspects when designing a system. However, an integrated approach is absolutely imperative. Furthermore, we believe that even persons who are not completely convinced of this thesis will eventually profit. The analogy with software development is quite apt. Software has been developed for over three decades; but only theoretical advances and design methodologies resulted in a significant improvement of the quality of software, the

benefits of which we are only beginning to see today. Note that this applies even to software written by people or teams who do *not* strictly adhere to those methodologies. We expect the same phenomenon with data security. From this conviction stems our belief that a good deal of the seemingly theoretical material in this text is in fact very practical.

This book is addressed to more than one type of reader. In the university or teaching setting, it is appropriate for a graduate course as well as an advanced senior course in computer science or management information systems. In the industrial setting, it should be of interest to professionals who are dealing extensively with data which very frequently are proprietary, such are systems analysts and systems or database designers and managers. Finally, the book should also appeal to researchers who want to gain a comprehensive perspective of the field.

An effort was made to keep this book completely self-contained. We think we succeeded, except perhaps for one or two areas, where references are provided for readers who would like to get more information than is needed for the purpose of this book. We do, however, assume the usual level of mathematical maturity one might expect from a person with a Bachelor's degree in the sciences or in engineering. The material presented is based on lectures, namely, a summer course for advanced graduate students at the Universidad de Chile, Santiago, taught in 1979, and a regular graduate course at the University of Houston, taught in 1980 as well as in 1981.

The book is organized into three major parts, which are preceded by an introduction. The first part deals with the security of statistical databases, the second part discusses authorization systems, and the third part covers the fundamental ideas of cryptography. Despite the general presentation, heavy emphasis is on methods which are practically useful. The text contains a number of (not too difficult) exercises which form an integral part of the material; thus the reader should at least attempt a number of them. The book also contains a fairly extensive bibliography.

Ernst L. Leiss

Contents

Chapter 1

Data Security: Introduction and Motivation

In the last few years, the amount of information stored in electronic media has increased significantly. Information which is stored in some medium is usually called *data*. An important aspect of the storage of information is the way in which access is performed; this is commonly called *information retrieval*. Here, practically always some knowledge is required in order to be able to access the information. This knowledge can come in many different forms. For example, in order to be able to withdraw money from an automated teller, a certain code must be supplied. This code is assigned to the account holder personally and is to be kept confidential. It should be noted that this is indeed an example of limiting access to information, as it is the stored information ("How much money is left in the account?") which is ultimately translated into more tangible resources (i.e., bank notes). Of crucial importance is the requirement that without this knowledge the stored information cannot be accessed. Clearly, it would be highly undesirable if someone other than the account holder, i.e., someone who does *not* know the code, could withdraw money from the account. Since this vital knowledge is deliberately hidden, presumably by the owner of the resource, access becomes impossible for anybody but the original owner. Thus the resource is protected, the information is secret. Clearly, the owner can pass on this knowledge to selected persons who then also have access. For example, the code which enables one to withdraw money from an automated teller can be given to other persons; this act will confer the authority to withdraw money from a specific account. However, some-

times someone not in such a group is able to acquire the necessary knowledge in one way or another. Obviously this is undesirable and should be avoided if at all possible. This problem leads naturally to the question how well protected, in other words, how *secure*, the resource is.

Security is a measure of the *effectiveness* with which the resource is protected. For the purpose of this book it describes how difficult it is for an unauthorized person to obtain the knowledge required to access the confidential data. For example, if an account holder's code for the automated teller is easily obtainable (e.g., if it is printed on the checks), access to the account is virtually not protected, the resource is clearly not secure.

The desire to maintain certain information secret is nothing new; indeed it is probably as old as any method to store information. The fundamental question involved is the right to access data (and, through the data, resources). However, there is an important difference between physical resources and information, although access to physical resources *is* very often governed by access to information. A thing, say a car, is of use only to that person who possesses it; it exists uniquely. Information, on the other hand, can be used by anybody who knows, i.e., has access to it; it can be duplicated without losing its usefulness. Consider for example the exchange of funds between financial institutions. In earlier days, a physical object, money or gold, was exchanged. This object could be stolen but not duplicated. (We disregard the forging of bank notes, as forging is *not* perfect duplication.) Nowadays, this exchange of funds is achieved by exchanging *information*. It is conceivable that this information is intercepted and slightly modified (for example, the recipient could be changed). Then the modification as well as the original information are transmitted. The delay introduced into this transmission could be imperceptible to the receiving institution; furthermore both messages will appear genuine. This example demonstrates that additional safeguards are necessary to prevent the duplication of information which may allow unauthorized access to resources (in this case, funds). Thus, while it is sufficient to possess a thing to prevent anybody from using it, the situation in the case of information is entirely different; it must be kept secret to achieve this objective and it should be as secure as possible in order to maintain this inaccessibility to unauthorized persons.

While the desire for secrecy and security of data is quite old, recent developments in data processing and information retrieval mandate a reevaluation of existing methods for achieving these objectives. Only

on the surface is this a question of quantity rather than quality. In fact, this (previously unheard of) amount of information does create *new* problems. To see this let us use an example which admittedly somewhat dramatizes the situation but nevertheless correctly illustrates the important point. The basic quality of a weapon is that it destroys. Thus the difference between a knife and a hydrogen bomb is only one of quantity: The bomb destroys more. However, in a wider context there are qualitative differences which are self-evident. A similar situation exists in the area of data security. Before the advent of high-speed computers the amount of data stored was rather limited, the information was stored in one place only, access was possible only to persons who were physically at the same place, it was laborious and time consuming, and finally it usually was possible in one way only (e.g., an office in a village which registered births and deaths when asked for the names of the men who are fathers of more than three but less than eight children who are still alive would be forced to sequentially search all the files several times, clearly a prohibitively time-consuming process). As a consequence, less information was asked to be supplied (because less could be handled) and security was much larger (as the information was highly distributed, retrieval required the physical presence of the person who wanted to obtain the information, and many queries were simply impossible or infeasible). A third group of observations pertaining to data security before the advent of computers can be summarized by the terms *smaller administrative units with little interaction* and *less mobility*; in other words, as people were born, worked, married, had children, and died more or less in the same place there was little need for an established way of information exchange. Furthermore the little exchange which did take place usually took on a very personal form.

Most of these (data-security-enhancing) restrictions on the interchange of information have been removed: Huge amounts of information are requested and stored. Often its relevance can only be guessed at, but the suspicion remains that at some point it could be used against its supplier. Access is almost unlimited; from virtually anywhere in the world almost anybody can pose a practically unlimited number and variety of queries in order to retrieve information from a vast diversity of sources. All that is required is a computer terminal as well as a telephone, in addition to the necessary authorization to use these resources (passwords, keys, etc.). Oftentimes, the same information is stored in more than one place, which, among other problems, creates the one that correction of wrong information in one place does not guarantee

correction of all instances of this incorrect information (credit reports!). Interchange of information is all pervasive, mandated by the ever-increasing size of administrative units and by the mobility of people.

The objective of this book is limited to a discussion of technical solutions to problems inherent in data security. The reader should, however, be aware that many problems require other approaches (such as legislation). Our scenario is the following. Suppose we are given a collection of confidential data which we want to keep secret. However, the nature of the data requires that certain restricted access be granted, restricted either in terms of the group of persons who are authorized to use the data or in the sense that only information based on certain aspects of the data be divulged but not individual data items themselves. We are interested in a precise description of the kind of access which can be allowed without violating the confidentiality of the information. Our point of view is that of an end user in a database–data communication system; in particular, we will not worry about the external, i.e., physical, security of our system. Similarly, we will assume that the operating system performs as it is supposed to perform. Specifically, we will assume that the operating system is secure, i.e., does not contain features which allow circumvention of protection mechanisms. In view of recent results on secure operating systems this is not an entirely unreasonable assumption. (Readers who consider this assumption too controversial and unrealistic are reminded that only by verifying the correctness of one module at a time, the correctness of a large system can be demonstated.) Below follows a very brief summary of the material in the next three chapters including a justification of its selection at the expense of the exclusion of other topics.

We distinguish three main areas where security is of great importance and which are of relevance to a user of a given system. The first area is that of databases, in particular, statistical databases, where a user might be interested either in the question whether a given database with given access mechanism is secure (i.e., "Do I want to use this database to make my confidential collection of data available on a restricted basis?") or in the problem of designing his-her own database which satisfies certain requirements concerning its security under certain access mechanisms (i.e., "What kind of database do I have to design if I want this and that?"). The important observation here is that everybody should probably be allowed to extract information based on certain aspects of the data but should not be able to determine any particular data item.

The second area is that of authorization. The scenario here is as follows. A user has a data object which is to be accessible only to a restricted group of people, i.e., each of these persons is authorized to access it in a certain way. The kind of authorizations and the way in which they are processed is almost always beyond the user's influence. However, the owner of such a confidential data object is probably very interested in the question of whether a certain authorization which s/he intends to grant has an adverse effect on the overall security (i.e., "If I give A this right and B that right will that enable C to do something terrible to me?").

The third area, finally, is that of encryption. It can be considered another aspect of the previous problem. If the owner of a confidential data object allows a certain restricted group to access this object, secret information is transmitted. These transmitted messages could be intercepted by an outsider; thus the secrecy of the data object would not be maintained. A well-known scheme designed to avoid this is to encode the information which is to be transmitted in such a way that only the intended receiver is able to decode it, i.e., while (physical) interception is still possible, the intercepted message is useless as it cannot be understood by the intercepter (i.e., "Can I send you this confidential message without somebody else eavesdropping?").

All three areas we propose to investigate have in common that they are of very practical importance to anybody who owns confidential information which is used (obviously, if it is not used it can simply be locked up, in which case there are no problems about data security), and that to a large extent they are under the user's control. The chapter on authorization mechanisms (Chapter 3) does deal extensively with a problem which is of importance in operating systems, namely, the safety problem; however, we present it in a somewhat more general context and without making explicit reference to operating systems. Finally, another very important reason for choosing these three areas is the fact that within the last few years a considerable body of knowledge related to them has been accumulated and that in all three areas there are now practical and useful methods available to resolve the problems addressed here in a realistic way.

Bibliographic Note

A discussion of the questions of statistical database security, authorization mechanisms, and en- and decryption can be found in Denning

and Denning[23] as well as a survey of other problems outside of the scope of this book, such as design of secure operation systems. This paper also contains an extensive bibliography.

Chapter 2

Statistical Database Security

2.1. Introduction

2.1.1. Motivation

A statistical database is a collection of data items each of which is to be considered confidential but access to statistical information about these data items is to be allowed. The central questions of this chapter can be formulated as follows: Is a given statistical database secure, i.e., using the responses to permitted queries is it possible to determine the previously unknown value of any of the confidential data items stored in the database, and more fundamentally, are there secure statistical databases?

At first this question appears strange. Since we do not allow access to individual data elements, it is intuitively clear that the confidentiality of our data is maintained, despite the fact that access to statistics is permitted. However, it turns out that our intuition is misleading. In this section we try to convince the reader informally that this intuitive feeling is by no means correct, in fact that there is a very definite problem in protecting the security of even the most restrictive statistical database. In subsequent sections we will give formal verifications of our assertions.

Consider a database containing grades of students who took a certain course, say, freshman composition. Each record has five keys, namely, SEX (N for female, Y for male), MEMBER of a fraternity/sorority (N for no, Y for yes), CLASS (from 1940 to 1979), OWNER

of a car (N for no, Y for yes), and WORKING (N for no, Y for yes). We assume that each possible sequence of five appropriate symbols identifies at most one student; however, given such a sequence there need not exist a student satisfying the description. Each record has two other fields, one containing the NAME of the students, the other his/her GRADE (between 0 and 100). For the purpose of this example, the grade will be considered confidential information. Now suppose that the owner of this database allows a sociologist to obtain statistical information based on the data stored in the database. The security of the database is to be enforced in such a way that only queries which involve at least two entries of the database are permitted, in which case the response is the sum of the grades of all students involved in the query; otherwise the response is "not permitted query." For example, the sociologist might pose the query (*, N, 1977, Y, Y), i.e., the sum of the grades of all the students who were not in a fraternity or sorority, took the course in 1977, owned a car, and worked. The asterisk indicates that the students can be of either sex. If there are two students the response will be the sum of their grades; if there are fewer than two the response will be "not permitted query." Note that there cannot be more than two, one female and one male, but there could well be none. Our sociologist, who supposedly is unable to access individual records, now poses the following four queries and receives the corresponding responses:

q_1 (*, Y, 1978, N, Y) : 142

q_2 (*, Y, 1978, *, Y) : 206

q_3 (Y, Y, 1978, *, Y) : 134

q_4 (*, Y, 1978, Y, Y) : "not permitted query"

We claim that these four queries suffice to determine several individual grades. First we see that the record (N, Y, 1978, Y, Y) does not exist because (Y, Y, 1978, *, Y) is permitted, thus (Y, Y, 1978, N, Y) and (Y, Y, 1978, Y, Y) must exist but (*, Y, 1978, Y, Y) is not permitted. For brevity denote

(N, Y, 1978, N, Y) by x_1

(Y, Y, 1978, N, Y) by x_2

(Y, Y, 1978, Y, Y) by x_3

All three records exist. We now can rewrite the queries q_1, q_2, q_3 as a

system of linear equations:

$$\begin{aligned} x_1 + x_2 &= 142 \quad (\text{from } q_1) \\ x_2 + x_3 &= 134 \quad (\text{from } q_3) \\ x_1 + x_2 + x_3 &= 206 \quad (\text{from } q_2) \end{aligned}$$

and this is equivalent to

$$\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 142 \\ 134 \\ 206 \end{bmatrix}$$

Since this matrix is nonsingular we can solve the equations, and this yields

$$\begin{aligned} x_1 &= 72 \\ x_2 &= 70 \\ x_3 &= 64 \end{aligned}$$

Hence we know that the students identified by (N, Y, 1978, N, Y) obtained 72, (Y, Y, 1978, N, Y) obtained 70, and (Y, Y, 1978, Y, Y) 64. The reader should realize that nothing was known beforehand about the database other than that an existing key sequence determines a unique element. Furthermore the intuitively very strong requirement that queries involving less than two elements will not be answered at all fails miserably and with it our intuition. In order to convince the reader that we did not cheat we reproduce the database for 1977 and 1978 in Table 2.1.

Table 2.1. A Simple Database

NAME	SEX	MEMBER	CLASS	OWNER	WORK	GRADE
A	N	Y	77	Y	N	74
B	Y	Y	78	N	Y	70
C	N	N	77	N	Y	86
D	Y	N	77	Y	Y	78
E	N	N	78	Y	N	67
F	Y	Y	78	Y	Y	64
G	N	Y	78	N	Y	72
H	Y	N	77	Y	N	69
I	N	N	77	Y	Y	73
J	Y	Y	77	Y	Y	60

While it is clear that this example demonstrates the failure of our intuition, it is not clear *why* it works. In order to provide insight into this "why" we will have to revert to a more formal presentation of the question of data security. We do hope, however, that this simple example provides sufficient justification for the more and more increasing concern about security of statistical databases as schemes similar to this have been (and are still being) used extensively in "real-world" databases.

2.1.2. Models and Methodology

In this section we first describe two models of statistical databases which are distinguished by the form of their queries. Both are simplifications of real-world databases, but reflect the main properties which are important from the point of view of security. Then we define formally what we mean by the security of such databases.

First we introduce the characteristic-specified model (DC). A database in this case is a partial function DC from keys of length K into the real numbers R, where K is fixed throughout and at least 2, i.e.,

$$\mathrm{DC} : \{0, 1\}^K \to R$$

For instance, for $K = 2$, $\{0, 1\}^2 = \{00, 01, 10, 11\}$ and DC assigns to some or all elements a real number. If DC is a total function then $\mathrm{DC}(w)$ exists for all keys w in $\{0, 1\}^K$ otherwise there are some keys v for which $\mathrm{DC}(v)$ is not defined. The range of DC is the confidential information of the database we will be concerned with. Characteristic-specified queries of type f are strings from $\{0, 1, *\}^K$ with the following interpretation:

(a) f is a function of arbitrarily many arguments (e.g., average, median, maximum, minimum, etc.).

(b) If the database manager is presented with a query q in $\{0, 1, *\}^K$ of type f it first determines all keys w in $\{0, 1\}^K$ which are matched by q and then applies f to the values $\mathrm{DC}(w)$.

A key w is matched by a query q if for all $i = 1, \ldots, K$ either the ith positions of w and q are identical or the ith position of q is an asterisk (*). Since the information in such a database can obviously not be kept confidential if queries involving only one element are allowed we will require that a query return a value only if it matches at least two existing keys. (Generalizations to more than two keys are obvious.) In particular, if DC is a total function this is the same as saying that q in $\{0, 1, *\}^K$ must contain at least one asterisk. Clearly for partial functions this is