LBL-30316
UC-400

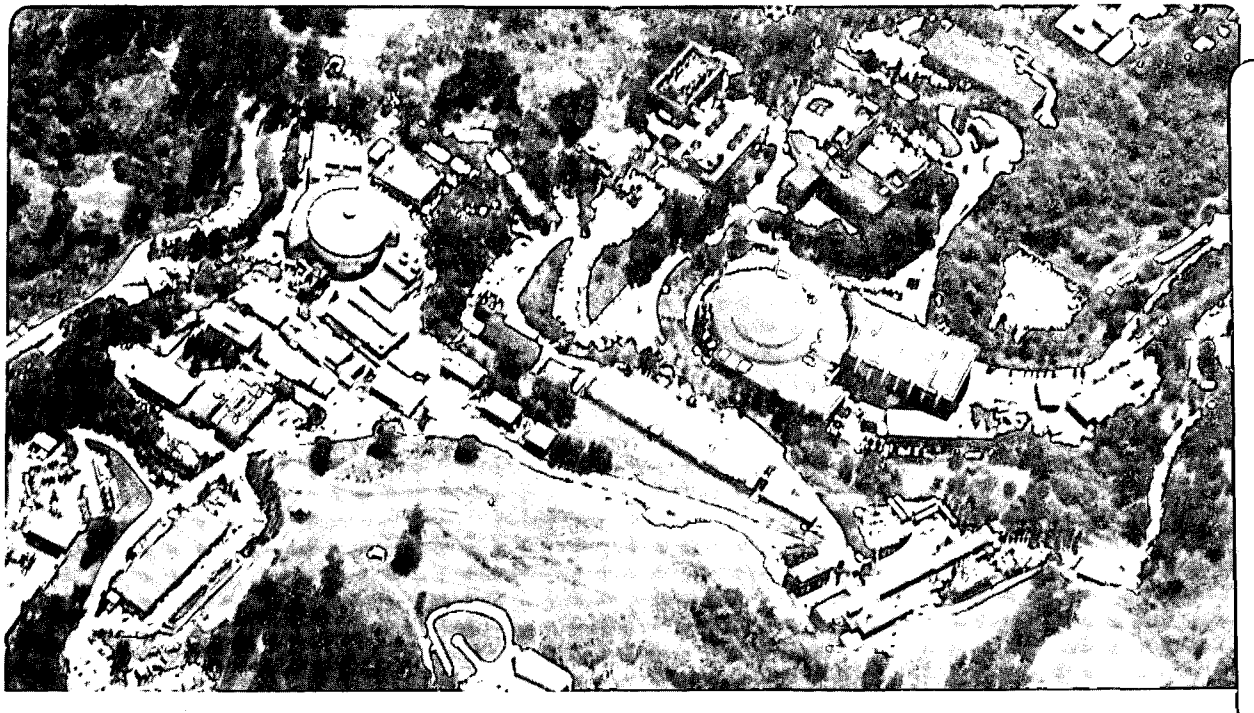
Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

Information and Computing Sciences Division

Security Control Methods for CEDR

D. Rotem

September 1990



Prepared for the U.S. Department of Energy under Contract Number DE-AC03-76SF00098

DISCLAIMER

LBL-30316

Security Control Methods for CEDR

Doron Rotem
Information & Computing Sciences Division
Lawrence Berkeley Laboratory
University of California
Berkeley, CA 94720

September 1990

This work was supported by the Epidemiologic Studies Division of the Office of Epidemiology and Health Surveillance, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

. Security Control Methods for CEDR

1. Introduction

1.1. Overview

In order for CEDR to be useful for analysts it must contain sensitive data about individuals such as medical information, job history, and educational background about individuals. On the other hand, there is a requirement that users of CEDR should not be allowed to infer confidential information about any specific individual represented in the database. This requirement appears in the SPEERA report as the following quotations show:

Section 5.4 - "Confidentiality protection must be provided for any data which contains personal identifiers"

Section 5.6- "All researchers should have full access to a basic health data set that has full protection against the identification of individuals"

"The Department should establish such a database and set procedures for public access"

This conflict, between the individual's right to privacy and society's need to know and process information, is by no means specific to CEDR and has been the subject of research studies in the area of inference controls on statistical databases. The objective of these studies is to find ways to provide users of the database with aggregate statistics about the collection of entities represented in it while protecting the confidentiality of any individual entity.

The purpose of this document is to summarize the findings of recent studies on the security problem in statistical databases and examine their applicability to the

specific needs of CEDR.

1.2. Organization

The document is organized as follows: In Section 2 we describe some general control methods which are available on most commercial database software. In Section 3 we provide a classification of statistical security methods. In Section 4 we analyze the type of users of CEDR and the security control methods which may be applied to each type. In Section 5 we summarize the findings of this study and recommend possible solutions.

2. Administrative and Software based controls

2.1. Administrative Controls.

Any technical solution to the security problem must be complemented by a system of administrative controls. Such controls include procedures by which users are given specific privileges with respect to the way they may use the database. In some cases such controls may allow only indirect access to the information or may require users to submit study proposals before any access privileges are granted.

In the following sections we will use the term DBA (database administrator) to refer to the person(s) whose duty is to implement and enforce both the administrative and technical controls.

2.2. Software Controls

Most database management software provide some built in mechanisms which can be used for statistical security controls. We list some of them in this section:

Authorization Each user who wishes to access the database will be identified by the system. Authorization tables will be maintained by the system to check permissions

on each operation attempted by the user. Auditing- This is a feature which may be used to monitor user activity within a database system. For example, the DBA can monitor who used the database and which tables were accessed. He can also monitor selective queries and record these results in an audit trail.

Permissions- Part of the standard query language are statements such as GRANT and REVOKE which allow the DBA to assign permissions to users to perform operations on tables such as access a table, update a table etc. For example, given a table called WORKER in the database

WORKER(BIRTH_DATE,STATE_BIRTH,SEX,RACE,SOCIO_ECONOMIC_STATUS,EDUCATION_LE

the statements:

GRANT SELECT, UPDATE ON WORKERS TO USER1; GRANT SELECT ON WORKERS TO USER2;

Allows USER1 to retrieve and change values in the table WORKER but USER2 may only do retrieval operations on this table.

Views- A DBA can control what portion of the database is visible to each user by creating views which are virtual tables. Views may include portions of base tables or new tables which are generated by some combination or manipulation of base tables. The GRANT and REVOKE operation mentioned above can be applied to views. Views can be created using a standard query language such as SQL. For example we may decide to create a view WORKER1 which hides some of the information in the WORKER table mentioned above. The following SQL statement achieves that:

```
CREATE VIEW WORKER1
AS
SELECT STATE_BIRTH,SEX,EDUCATION_LEVEL
FROM WORKER;
```

Encryption- Special fields such as names and numbers used for identification can be stored in encrypted form so that only authorized users can see them.

3. Statistical Security Control Methods

3.1. Motivation

The methods mentioned in the previous section are not enough to stop a sophisticated attempt to compromise the database which uses a sequence of queries for inference. This imposes some challenging requirements on the design of the query interface since a user may attempt to infer confidential information by correlating a sufficient number of summary answers. The user may also base his/her inference on some additional information available from external sources which is not explicitly stored in the database. As an example let us consider a database which holds information about AGE, SEX, EMPLOYER, and DIAGNOSIS of individuals. Let us assume that the value of DIAGNOSIS is confidential. Let P1 be the predicate:

P1: (AGE=42 and SEX= Male and EMPLOYER =ABC)

and suppose a counting query Q1

Q1: COUNT (P1)

returns the answer 1. We can then follow by second counting query Q2 where:

Q2: COUNT(P1 and DIAGNOSIS = Lung Cancer)

In this case if the answer to Q2 is also 1, we establish a diagnosis for an individual, if it is 0 we can keep on trying other possibilities until we find the right one.

Even if we refuse to answer queries whose count is 1, the database can be compromised by the sequence:

Q1 : COUNT (SEX =FEMALE)

answer 10053

Q2 : COUNT (SEX=FEMALE or P1)

answer 10054

Q3 : COUNT (SEX=FEMALE or (P1 and DIAGNOSIS=Lung Cancer))

any answer here will give us some information about an individual.

It is therefore important to ensure that no sequence of queries is sufficient to deduce private or confidential information about any individual. This task is not simple and requires query filters as well as some query logs which register which previous queries were submitted by users. In this section we survey some methods which provide security against such sophisticated attempts.

3.2. Modeling

As the previous section shows, in order to understand all the aspects of a statistical security system we need to model the following:

• Object Definition Construct

A statistical database may contain data about different objects, i.e., workers, facilities, treatments etc. The security control mechanisms should take into account the security requirements of each of these objects and their subcategories.

• User Knowledge Construct

We need to keep track of the knowledge properties of each group of users. These propertiers include existing knowledge and also knowledge gained by a chain of previous queries.

• Constraint Enforcer and Checker

This is the process which intercepts queries which are submitted to the database and enforces the security mechanisms associated with them. In case the query is performed successfully, it also updates the information the system maintains about user knowledge.

Maintaining the above constructs seems to be a very difficult task and to the best of our knowledge has not been implemented in any existing system. However, they provide a framework against which current systems can be compared and measured.

3.3. Classification of Methods

It is agreed by most researchers in the field that privacy protection schemes are acceptable only if they are both effective and efficient. A method is effective if it provides statistical security to the maximum extent possible while maintaining the richness of information available to the users. It is efficient if it does not cause significant performance degradation.

There are many protection policies suggested in the literature which claim to satisfy the above constraints. In general they can be classified as follows:

- Query Restriction - This is a family of methods which limit the type of queries allowed. The main methods here are: query-set-size control, query overlap control, auditing, cell suppression and database partitioning.
- Perturbing the database- Changing the values of the real database by small variations and making only the perturbed database available to users. For example, to avoid some possibility of inference about individuals, we may maintain information such as date of birth, job classification years of education etc. in approximate form, e.g., instead of complete date of birth month-day-year we can store only year. Other possibilities include rounding and maintaining ranges for each value rather than the value itself. Encryption can also be viewed as a perturbation method although no loss in precision is incurred in this case. All perturbation methods must be implemented carefully as to ensure that the data is precise enough for analysis purposes.
- Perturbing outputs- Changing the answers slightly without changing the actual database. All database perturbation methods mentioned above, can also be used only for output perturbation. For example it can be shown that by adding or subtracting a small constant from answers which include counters, the chances of compromising the database reduce significantly without affecting statistical validity. Another possibility is to supply only results based on random samples

from the database rather than answers based on the population itself. For example if a user asks what is the average age of a group of workers based on some criterion, the database computes a random sample from that group and gives the sample average as an answer.

In most cases, a security control strategy combines one or more of the above methods.

In Table 1, we summarize the different methods and rank them based on the implementation efforts needed and the loss of precision to the database in case they are adopted.

3.4. The Lattice Model

Most of the methods of type query restriction can be explained within a framework which is commonly mentioned in the research literature called the **lattice model**. This method treats a database with k categorical attributes as a collection of logical tables of counts of dimension at most k. More specifically, the location of each cell in the table represents a combination of attribute values. A cell contains a count of the number of records in the database with attribute values corresponding to the cell location in the table. The user may be presented with a view of the database which is a collection of tables of dimensionality between 1 and k obtained by aggregation across one or more of the k dimensions.

Researchers in this area have shown that a collection of tables obtained from each other by aggregation form a well known algebraic structure called a lattice. By exploring properties of this lattice they are able to determine which collections of tables should not be presented together as they may cause a compromise of the database.

In a database defined over k attributes,the lattice is constructed as follows. At the bottom (level k), there is a single k dimensional table and at level 0 there is one table with a single cell counting the number of tuples in the database. In general, at level j

we have all possible j dimensional tables. Tables in two successive levels are connected if one can be obtained from the other by aggregating across one dimension. In other words, a table T at level j-1 is connected to a table M at level j if the counts in T can be computed from these in M by aggregating across one of the attributes of M.

For example, consider a database with k=6 which stores information about attributes of workers such as Birth Date (BD), State of Birth (SB), Sex (S), Race (R), Socio-Economic-Status (SES) and Education Level (EL). The general structure of a lattice for this example is shown in Figure 1.

Based on this model, it can be shown that additive statistics computed on a database are linear combinations of counts in various tables associated with a database. More formally, statistics are collected for subsets of records having common attribute values using logical formulas. A logical formula uses the operators AND, OR and NOT with simple predicates over attribute values. For example the formula Q,
Q = {(S= "F") AND (SB = "CALIFORNIA) OR (SB = "NEW_YORK)}
denotes the subset of all female employees born in California or New York.

A set of records specified by a formula Q is called the query set of Q. A query set is of dimension m if it can only be specified by a formula with m distinct attributes. The above formula is of dimension 2. The cardinality of the query set of any query of dimension m can be computed from the cells of an m dimensional table of counts.

For example, in the figures below we constructed a 2 dimensional table for the attributes SB (state of birth) and EL (education level) encoded as a number between 1 and 7. From this table we obtain two one dimensional tables by aggregating across each of the two attributes. From each such table we can obtain a table of one cell showing the cardinality of the database.

It can be shown that most inference problems can be modeled as special relationships between tables known to a user. Suppose that a decision is made that cells with

value 1 are sensitive and should not be revealed. A user can still infer about such cells if he is allowed to freely access cells which are not considered sensitive. For example if a user can access the table T_{EL}, and also knows the number of individuals in CA or MN with education level 6, he may infer the value of a sensitive cell (NY and EL equals 6) by simple arithmetic manipulations.

Most proposals for query restrictions can be interpreted in terms of the lattice model. For example, cell suppression is a restriction which does not allow to reveal the contents of cells whose count fall below some predefined threshold.(See Figure 3). Another method combines cells together to avoid cells with low counts. In Figure 4 this is shown where two education levels (6 and 7) are combined into a single column. Another example is that of substituting ranges instead of exact values as shown in Figure 5.

4. Implementation of security control methods in CEDR

Any database security strategy must take into account the nature of the confidential information which must be protected as well as system user profiles. Based on our interviews, we envision at least three types of users for CEDR:

1) Database Administrators: Users who are responsible for accepting new data to CEDR and enforce integrity controls in order to maintain it in a reliable state. In order to perform their tasks they will be allowed to insert, delete and modify any portion of the database.

2) Analysts: Researchers who perform epidemiological studies on the population stored in CEDR. These researchers will typically perform statistical analysis of the data. In addition, they need access to individual records in order to ensure consistency and accuracy of such records. They will be allowed to do so as long as individual identifiers are not revealed to them. In addition, such

users will also submit data for possible inclusion in the database.

3) Casual Users: These are users who propose to conduct studies on the database. They will be allowed to perform statistical studies and retrieve aggregate information but will not have permission to view any individual records. They will also be allowed to browse the metadata.

We assume that effective access control mechanisms will be implemented in CEDR to avoid unauthorized access to the system. Such controls include identification and authentication of each person seeking access to CEDR as well as his/her user type. Our main effort will therefore be directed at avoiding improper use of the database by authorized users who try to access or infer confidential information from data available to them according to their user type.

Most of our protection efforts will be directed against users of type 2 and 3. Such users will typically work in two stages: In stage 1 they extract some subset of records from the database and in stage 2 proceed to manipulate these records with a statistical package such as SAS or MOX. It is almost impossible to provide any protection in stage 2 and therefore we need to control the data extracted from the database in stage 1.

In Table 2 we summarize the user types, their access requirements and the type of controls which are appropriate for them.

5. Summary and recommendations

- Privacy protection policy consists of administrative and technical components.
- The administrative component includes a user classification policy and procedures for assigning database access privileges.
- The technical component is carried out by database administrators who are

responsible for database design, creation of views, control of query types etc.

- Privacy protection will incur costs in implementation efforts and performance degradation of the database
- Tradeoffs exist between the level of protection and the implementation effort and richness of data available to users.
- Identifying variables (e.g. dates, age) could be presented with minimum amount of accuracy to minimize the risk of identification.

5.1. Recommendations

Our recommendations are summarized in Table 3. The controls we suggest are based on user types and also take into account the cost of implementation. Three user type categories are sufficient, a user will be classified by administrative procedures and authentication will be provided by the software. Under the assumption that loss in precision is unacceptable we eliminated all perturbation methods which lead to such loss from consideration. From the remaining methods we recommend using administrative controls to the maximum degree possible and technical controls which can be implemented at a low or moderate cost.

SECURITY CONTROL TYPE	METHODS	IMPLEMENTATION EFFORTS	LOSS OF PRECISION
QUERY RESTRICTION	QUERY SIZE CONTROL	LOW	NO
	QUERY OVERLAP CONTROL	HIGH	NO
	CELL SUPPRESSION	HIGH	NO
	DATABASE PARTITIONING	MODERATE	NO
DATABASE PERTURBATION	APPROXIMATE VALUES	LOW	YES
	ENCRYPTION	LOW	NO
OUTPUT PERTURBATION	RANDOM SAMPLE QUERIES	MODERATE	YES
	VARYING OUTPUT	MODERATE	YES
	ROUNDING	LOW	YES

TABLE 1

SECURITY CONTROL METHODS

USER TYPE	EXAMPLE	FUNCTIONS	ACCESS REQUIREMENTS	CONTROLS
1	ADMINISTRATOR DATA VALIDATOR	CREATE AND MAINTAIN DATABASE CHECK ACCURACY	ALL INFORMATION	NONE
2	ANALYST	PERFORM STUDIES CHECK ACCURACY SUBMIT DATA FOR INCLUSION	ALL INFORMATION EXCLUDING IDENTIFIERS	ACCESS ADMINISTRATION
3	CASUAL USER POTENTIAL ANALYST	PRELIMINARY STUDIES BROWSING	AGGREGATE DATA METADATA NO INDIVIDUAL INFORMATION	STATISTICAL ADMINISTRATIVE RESTRICTION OF VARIABLES

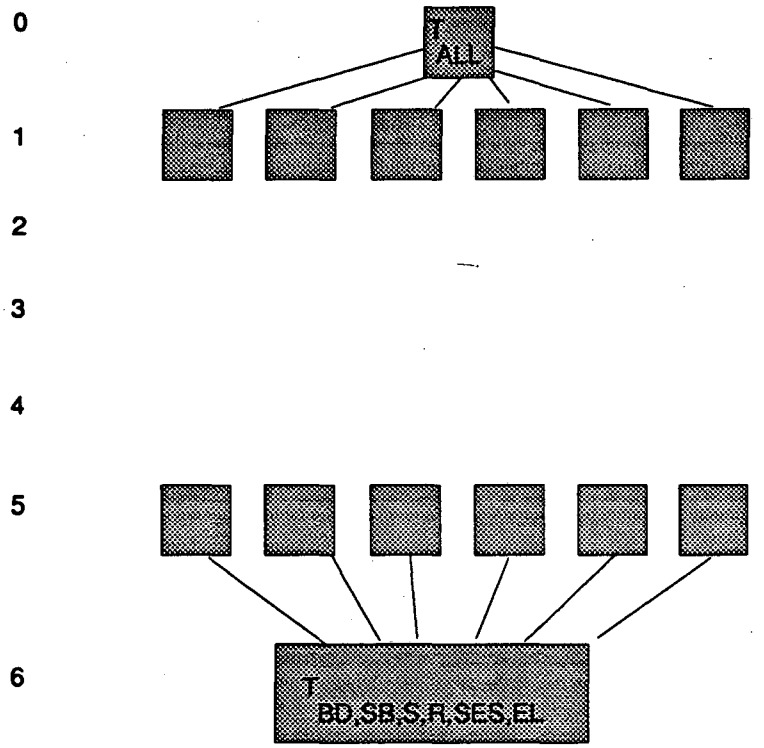
CEDR USER TYPES

TABLE 2

RECOMMENDATIONS

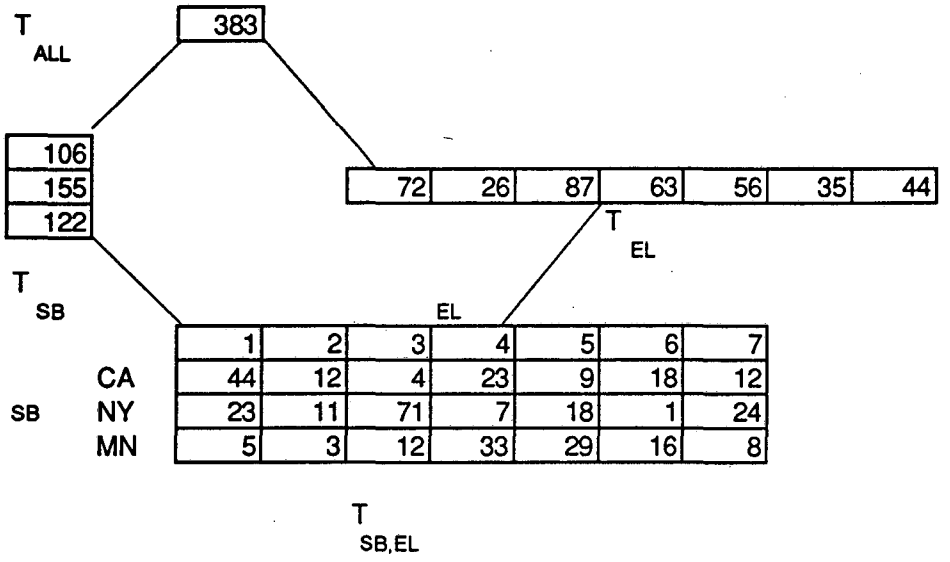
USER TYPE	RECOMMENDED MEASURE	COST
1	NONE	
2	1. JUST IDENTIFY TO THE MAXIMUM DEGREE (IS RISK ACCEPTABLE?)	LOW
	2. ADMINISTRATIVE ACCESS CONTROL	HIGH
3	1. FAKE DATA (PUBLIC USE DATA)	LOW
	2. GENERAL CONTROLS	
	VIEWS (PERFORMANCE DEGRADATION)	LOW
	ENCRYPTION(MORE RESTRICTIVE)	LOW
	3. STATISTICAL DATABASE PARTITIONING & CELL SUPPRESSION	MODERATE

TABLE 3



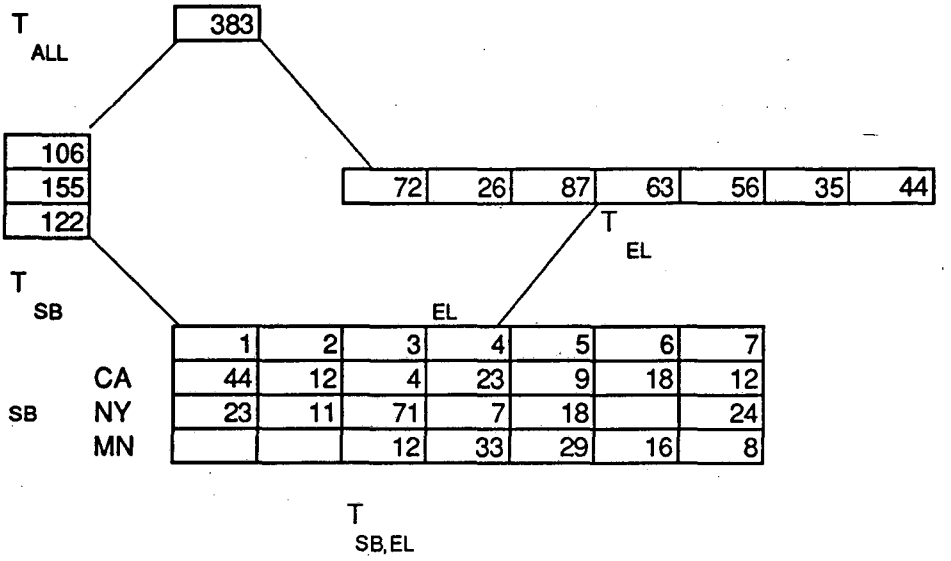
THE LATTICE MODEL FOR A DATABASE OF DIMENSION 6

FIGURE 1



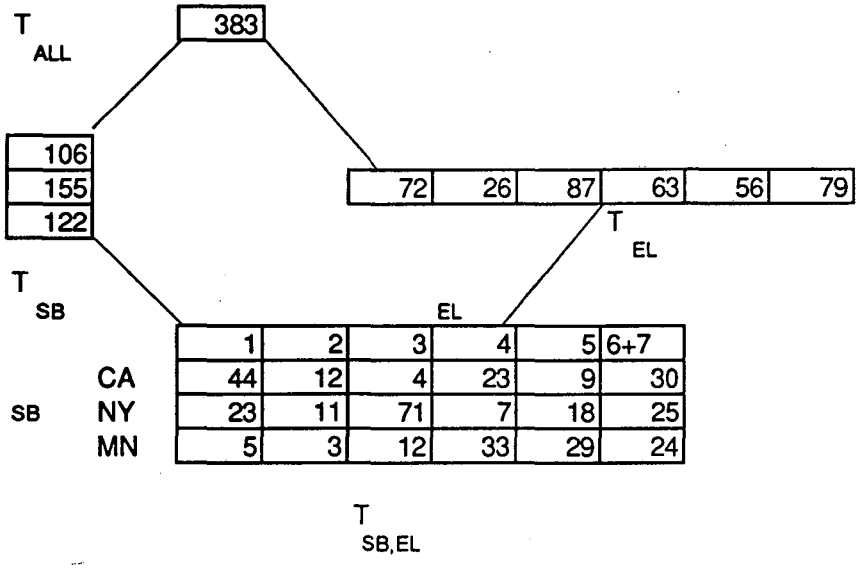
LATTICE OF TABLES FOR ATTRIBUTES SB AND EL

FIGURE 2



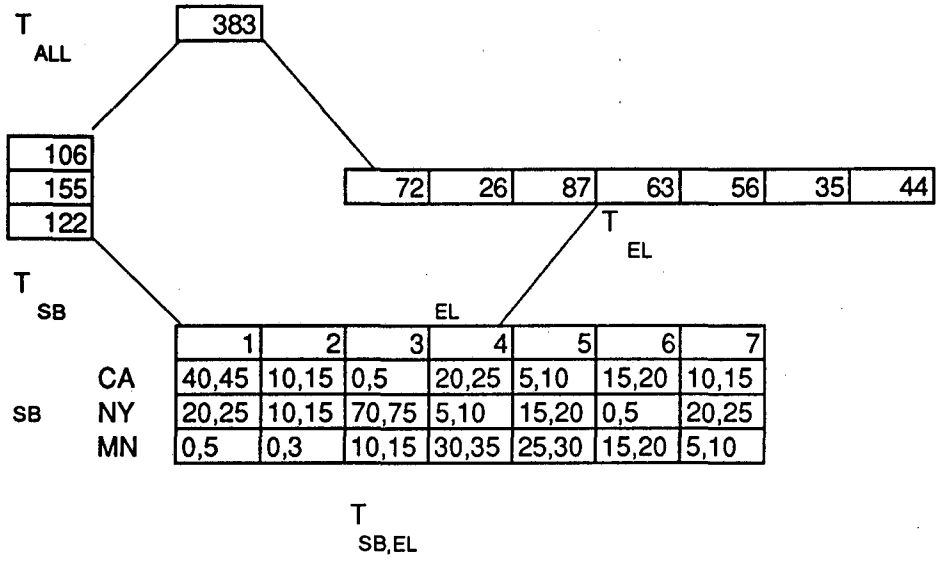
CELL SUPPRESSION OF CELLS WITH VALUES LESS THAN 6

FIGURE 3



GROUPING OF EDUCATION LEVEL 6 AND 7

FIGURE 4



SUBSTITUTION OF VALUES BY RANGES

FIGURE 5

LAWRENCE BERKELEY LABORATORY
UNIVERSITY OF CALIFORNIA
INFORMATION RESOURCES DEPARTMENT
BERKELEY, CALIFORNIA 94720