The Tracker: A Threat to Statistical Database Security

DOROTHY E. DENNING and PETER J. DENNING
Purdue University
and
MAYER D. SCHWARTZ
Tektronix, Inc.

The query programs of certain databases report raw statistics for query sets, which are groups of records specified implicitly by a characteristic formula. The raw statistics include query set size and sums of powers of values in the query set. Many users and designers believe that the individual records will remain confidential as long as query programs refuse to report the statistics of query sets which are too small. It is shown that the compromise of small query sets can in fact almost always be accomplished with the help of characteristic formulas called trackers. Schlörer's individual tracker is reviewed; it is derived from known characteristics of a given individual and permits deducing additional characteristics he may have. The general tracker is introduced: It permits calculating statistics for arbitrary query sets, without requiring preknowledge of anything in the database. General trackers always exist if there are enough distinguishable classes of individuals in the database, in which case the trackers have a simple form. Almost all databases have a general tracker, and general trackers are almost always easy to find. Security is not guaranteed by the lack of a general tracker.

Key Words and Phrases: confidentiality, database security, data security, secure query functions, statistical database, tracker
CR Categories: 3.7

1. INTRODUCTION

Statistical databases must supply statistical summaries about a population without revealing particulars about any one individual. Yet, statistical summaries contain vestiges of the original information: A questioner may be able to deduce the original information by processing the summaries. When this happens, the personal records are compromised.

Database designers and users would like to know when compromise is possible and, if so, how easy it is. We studied these questions in the context of databases having these properties:

This work was supported in part by the National Science Foundation under Grant MCS77-04835 at Purdue University.
Authors' addresses: D.E. Denning and P.J. Denning, Computer Sciences Department, Purdue University, West Lafayette, IN 47907; M.D. Schwartz, Tektronix, Inc., P.O. Box 500, Beaverton, OR 97077.

—Each individual's record is identified by a set of characteristics and contains one or more confidential values.

—A query program examines a "query set"— the collection of records whose characteristics match those of a given "characteristic formula."

A query computes a raw statistic for the query set, usually the sum of powers of values in records of the query set. Most statistical databases have these properties, and so do relational systems such as INGRES [20] or System R [1, 2].

Our point of departure is Schlörer's work, which showed that statistical databases can be easily compromised even if some queries are not answerable because their query sets (or complements) are too small [14]. The questioner divides his preknowledge of a given individual into parts, which are then reassembled into a special characteristic formula called a *tracker*. From the responses of a few answerable queries involving the tracker, the questioner may determine whether or not the given individual has a characteristic previously unknown to the questioner.

This paper continues the investigation of compromises based on trackers. There are four principal results. First, we will remove the dependency of the tracker on a specific individual. The *general tracker* permits the questioner to answer arbitrary queries without any prior information about anyone in the database. Second, we will show that tracker compromises apply to any statistical query, not just counts. Third, we will give a simple structural condition that guarantees the existence of a general tracker and specifies its form. This condition also reveals that almost all databases have trackers. Fourth, finding a tracker is usually not difficult.

The conclusion is that statistical databases are almost always subject to compromise. Severe restrictions on allowable query set sizes will render the database useless as a source of statistical information but will not secure the confidential records.

Literature

Hoffman and Miller presented a simple algorithm for compromising databases using counting queries based on conjunctive characteristic formulas, i.e. logical ANDs of category-values [10]. Haq formalized and extended these ideas [9], and Palme showed that they work for summing queries as well [13]. Fellegi and Hansen independently studied methods of protecting individual records in Census files [5, 8]; these methods, which are based on restricting queries to statistical samples of the very large database, cannot be used in small or medium databases. Schlörer showed how a tracker can be used to deduce additional characteristics of a known person even if the query system gives no answer when the query set (or its complement) is too small [14]. Effective countermeasures, which are hard to find, make compromise more difficult by modifying the data or the answers in some unknown way [6, 15, 21]. Dobkin, Jones, and Lipton studied compromises using queries that calculate sums over fixed size query sets [4]; we extended these results to include arbitrary linear functions over fixed size query sets [18, 19]. Kam and Ullman studied compromises in databases wherein there is exactly one record for each possible combination of the basic category values that can appear in characteristic formulas [11]. Chin studied compromises in databases which provide counts and linear sums of query sets containing at least two records [3].

2. MODEL OF A STATISTICAL DATABASE

A statistical database contains records for some number n of individuals. Each record contains confidential *category* and *data fields*; at least two values exist for each such field. The category fields are used to identify and select records, while the data fields hold other information. The category fields need not be disjoint from the data fields. (There may also be a unique identifier field, which is neither category nor data; it is not employed by any statistical query.) No updates or deletions are made during a period when compromise is being attempted.

Each query for this database uses a *characteristic formula* C, which is an arbitrary logical formula using category-values as terms connected by operators AND ($\cdot$), OR (+), and NOT ($\bar{}$). (SEQUEL is an example of a query language permitting such formulas [2].) The set of records whose category fields match C is called the *query set* X_C. The family of queries considered here compute raw statistics of the form

$$q(C; j, m) = \sum_{i \in X_C} v_{ij}{}^m,$$

where v_{ij} is the value in data field j of record i, and m is an integer. When $m = 0$, the query simply returns the size of the query set $|X_C|$ for any j; we call this a *counting query* and denote it by COUNT(C). When $m = 1$, the query returns the sum of values in the jth data field for records in X_C; we call this a *summing query* and denote it by SUM(C; j). The mth moment of the data in X_C is calculated from $q(C; j, m)$/COUNT(C). We will use the simple notation $q(C)$ to stand for any query in this family (for arbitrary j and m).

Table I shows a database summarizing confidential information about employees in a hypothetical university's College of Mathematical Sciences. Each person is classified in four categories and has two data values. The possible category-values are as follows:

Sex:	*M, F*
Dept:	*CS, Math, Stat*
Position:	*Adm, Prof, Stu*
Salary:	\$$N$ K *Sal,* for $N = 0, 1, 2, \ldots$

The possible data-values are:

Salary (in \$K):	any integer ≥ 0
Contribution (in \$):	any integer ≥ 0

Examples of queries for this database, expressed formally and informally, are as follows:

Formal query	Answer	Informal statement
COUNT($M \cdot CS$)	3	Number of males in the CS Dept.
COUNT($F \cdot Prof \cdot (CS + Math)$)	2	Number of female professors in either the CS or Math Depts.
SUM($M + \overline{CS}$; *Sal*)	\$176K	Total of salaries among either males or NonCS personnel.
SUM(\$15K *Sal*; *Contr*)	\$150	Total of contributions by persons earning \$15K.

Table I. Database Containing Information on Employees and Their Political Contributions, for a Hypothetical University's College of Mathematical Sciences

		Categories			Data	
No.	Unique identifier	Sex	Dept	Position	Salary ($K)	Political contribution ($)
1	Adams	*M*	*CS*	*Prof*	20	50
2	Baker	*M*	*Math*	*Prof*	15	100
3	Cook	*F*	*Math*	*Prof*	25	200
4	Dodd	*F*	*CS*	*Prof*	15	50
5	Engel	*M*	*Stat*	*Prof*	18	0
6	Flynn	*F*	*Stat*	*Prof*	22	150
7	Grady	*M*	*CS*	*Adm*	10	20
8	Hayes	*M*	*Math*	*Prof*	18	500
9	Irons	*F*	*CS*	*Stu*	3	10
10	Jones	*M*	*Stat*	*Adm*	20	15
11	Knapp	*F*	*Math*	*Prof*	25	100
12	Lord	*M*	*CS*	*Stu*	3	0

Characteristic formulas can be extended to permit relations, for example,

$$\text{SUM}(Sal \leq \$15\text{K};\ Contr) = \$180.$$

Extended characteristic formulas are merely abbreviations for larger formulas; they do not change the nature of queries. For example,

$$\text{“}Sal \leq \$15\text{K}\text{”} = \text{“}\$1\text{K } Sal + \$2\text{K } Sal + \cdots + \$15\text{K } Sal.\text{”}$$

3. COMPROMISE

A compromise occurs when a questioner deduces, from the responses to one or more queries, confidential information of which he was previously unaware. The compromise is "positive" if the questioner deduces the value in a given category or data field of a given individual. The compromise is "negative" if the questioner deduces that a value is *not* in a given category or data field of a given individual. In Table I, for example, a questioner who learns that Baker contributed $100 has effected a positive compromise; but if he learns only that Baker did not contribute $200, he has effected a negative compromise. A database is *secure* if no compromise is possible.

It is well known that compromise is easy when query sets can be small or large compared to the size of the database [3, 10, 14, 15, 17]. Two examples illustrate.

Example 1. A questioner who knows that Dodd is a female CS professor poses two queries in Table I:

$$\text{COUNT}(F \cdot CS \cdot Prof) = 1$$

$$\text{COUNT}(F \cdot CS \cdot Prof \cdot \$15\text{K}Sal) = 1$$

These queries reveal Dodd's salary, because she is the only possible individual satisfying the characteristics of both queries. Were the response to the second

query 0, negative compromise would result, since the questioner would deduce then that her salary was not \$15K. ■

Example 2. Because COUNT($\bar{C}$) = n − COUNT(C), the compromise of Example 1 can also be achieved with large query sets. The questioner first determines n by posing a query with a tautology as the formula; for example, COUNT($Prof + \overline{Prof}$) = 12. He then poses COUNT($\overline{F \cdot CS \cdot Prof}$), the response to which is 11. The difference, 12 − 11, is the number of female CS professors. The questioner can determine this person's salary (\$15K) by subtracting the responses of two more queries:

$$\text{SUM}(Prof + \overline{Prof}; Sal) = \$194\text{K}, \qquad \text{SUM}(\overline{F \cdot CS \cdot Prof}; Sal) = \$179\text{K}. \quad ■$$

Example 1 illustrates why a lower bound, say k, must be imposed on the size of the smallest allowable query set. Example 2 illustrates that, by symmetry, an upper bound $n - k$ must be imposed on the size of the largest allowable query set. Using the symbol ⅋ to denote an unanswerable query, we redefine queries (for given j and m) thus:

$$q(C) = \begin{cases} \sum_{i \in X_C} v_{ij}{}^m, & k \le \text{COUNT}(C) \le n - k, \\ ⅋, & \text{otherwise.} \end{cases}$$

When $k = 0$ this is the same as our earlier definition. Note that $k \le n/2$ if any queries at all are to be answerable.

The following sections show that compromise is possible even for relatively large values of k. All the methods are based on "trackers," special characteristic formulas which can be used to calculate indirectly the values of unanswerable queries. We begin with Schlörer's individual tracker, then turn to the general tracker and the double (general) tracker.

4. THE INDIVIDUAL TRACKER

Schlörer [14] considered the following problem for counting queries which are answerable only for query set sizes in the range $[k, n - k]$, where $1 < k \le n/2$. The questioner knows from external sources that a given individual I, whose record is in the database, is uniquely characterized by the formula C. The questioner seeks to learn whether or not I also has characteristic a. Since COUNT($C \cdot a$) ≤ COUNT(C) = $1 < k$, the questioner cannot use the method of Example 1. Schlörer showed that, if the questioner can divide C in two parts, he may be able to calculate COUNT($C \cdot a$) from two answerable queries involving the parts. This result can be extended to work for any statistical query $q(C)$.

Suppose that the formula C believed to identify I can be decomposed into the product $C = A \cdot B$, such that COUNT($A \cdot \bar{B}$) and COUNT(A) are both answerable:

$$k \le \text{COUNT}(A \cdot \bar{B}) \le \text{COUNT}(A) \le n - k. \tag{1}$$

The formula $T = A \cdot \bar{B}$ is called the *individual tracker* (of I) because it helps the questioner "track down" additional characteristics of I. The method of compromise is summarized below.

INDIVIDUAL TRACKER COMPROMISE. *Let $C = A \cdot B$ be a formula identifying individual I, and suppose $T = A \cdot \bar{B}$ is I's tracker. With three answerable queries, calculate:*

$$\text{COUNT}(C) = \text{COUNT}(A) - \text{COUNT}(T), \tag{2}$$

$$\text{COUNT}(C \cdot a) = \text{COUNT}(T + A \cdot a) - \text{COUNT}(T). \tag{3}$$

If $\text{COUNT}(C \cdot a) = 0$, *I does not have characteristic a (negative compromise). If* $\text{COUNT}(C \cdot a) = \text{COUNT}(C)$, *I has characteristic a (positive compromise). If* $\text{COUNT}(C) = 1$, *arbitrary statistics about I can be computed from*

$$q(C) = q(A) - q(T). \tag{4}$$

PROOF. With the help of Figure 1, we see that eq. (4) holds, and that

$$q(C \cdot a) = q(T + A \cdot a) - q(T). \tag{5}$$

The queries $q(A)$ and $q(T)$ are assumed to be answerable (relation (1)). The query $q(T + A \cdot a)$ is also answerable because its query set contains X_T and is contained in X_A, both of which are assumed to be answerable. Therefore the queries used on the right-hand sides of these equations are all answerable; $q(C)$ and $q(C \cdot a)$ are thereby calculable. Equations (2) and (3) result when eqs. (4) and (5) are applied with counting queries. ■

When $\text{COUNT}(C) > 1$, it may happen that no compromise is possible; this will be illustrated below in Example 4. But when $\text{COUNT}(C) = 1$, we may apply eq. (4) to discover the statistics for the given individual I. Equation (3) is Schlörer's result [14]. When applied with summing queries, eq. (4) is Palme's result [13].

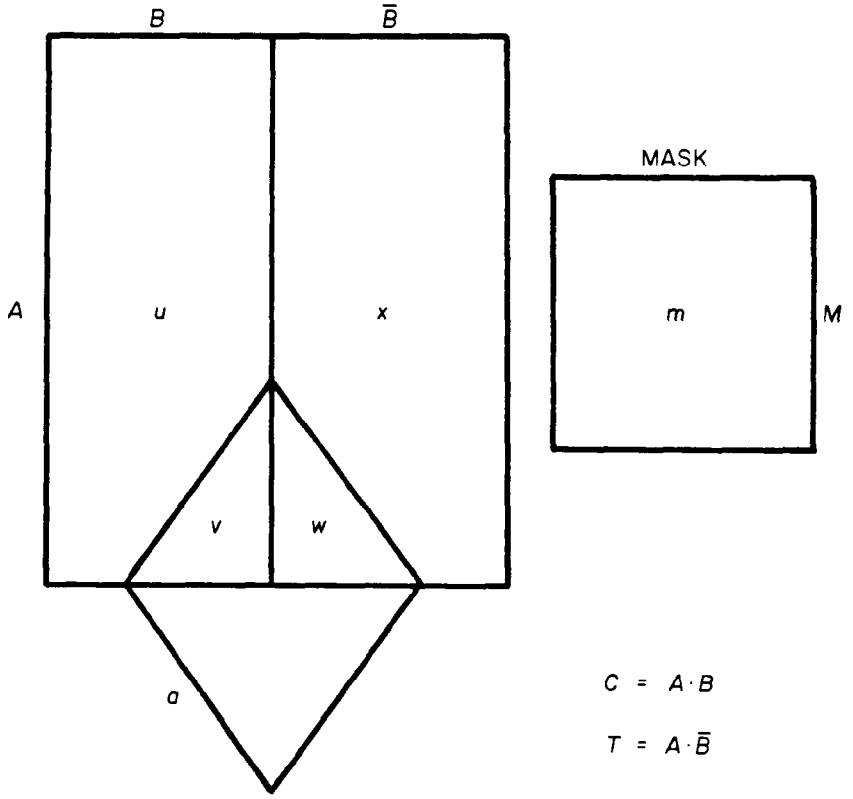
This compromise is not prevented by the lack of a decomposition of C giving answerable A and T. Schlörer pointed out that unanswerable formulas A and T can often be replaced with answerable $A + M$ and $T + M$, where $\text{COUNT}(A \cdot M) = 0$; see Figure 1. The formula M, called the "mask," serves only to pad the small query sets with enough (irrelevant) records to make them answerable.

Example 3. We will illustrate the individual tracker compromise for the database of Table I with $k = 2$. The query set size restriction implies that a query $q(C)$ is answerable only if $2 \leq \text{COUNT}(C) \leq 10$. A questioner believes that $C =$ "$F \cdot CS \cdot Prof$" characterizes Dodd, but the restriction $k = 2$ prevents his using the methods of Examples 1 and 2 to determine Dodd's salary. However, the questioner can make a tracker $T = A \cdot \bar{B}$ where $A =$ "F" and $B =$ "$CS \cdot Prof$." To verify that Dodd is the only individual characterized by C, the questioner applies eq. (2):

$$\begin{aligned}\text{COUNT}(F \cdot CS \cdot Prof) &= \text{COUNT}(F) - \text{COUNT}(F \cdot \overline{CS \cdot Prof}) \\ &= 5 - 4 \\ &= 1.\end{aligned}$$

To discover Dodd's salary by Schlörer's method, the questioner would have to search using repeated applications of eq. (3). If he guessed \$25K, eq. (3) would yield

$$\begin{aligned}\text{COUNT}(F \cdot CS \cdot Prof \cdot \$25\text{K}Sal) = \text{COUNT}(F \cdot \overline{CS \cdot Prof} + F \cdot \$25\text{K}Sal) \\ - \text{COUNT}(F \cdot \overline{CS \cdot Prof})\end{aligned}$$



WITHOUT MASK	WITH MASK
$q(A) = u+v+w+x$ $= (u+v)+(w+x)$ $= q(C)+q(T)$	$q(A+M) = u+v+w+x+m$ $= (u+v)+(w+x+m)$ $= q(C)+q(T+M)$
$q(T+A\cdot a) = v+w+x$ $= v+(w+x)$ $= q(C\cdot a)+q(T)$	$q((T+M)+(A+M)\cdot a) = v+w+x+m$ $= v+(w+x+m)$ $= q(C\cdot a)+q(T+M)$

Fig. 1. Venn diagram showing relations among queries used in the individual tracker compromise

$$= 4 - 4$$

$$= 0,$$

revealing that Dodd's salary cannot be \$25K. As soon as the questioner guesses \$15K, eq. (3) yields

$$\text{COUNT}(F \cdot CS \cdot Prof \cdot \$15\text{K}Sal) = \text{COUNT}(F \cdot \overline{CS \cdot Prof} + F \cdot \$15\text{K}Sal)$$

$$- \text{COUNT}(F \cdot \overline{CS \cdot Prof})$$
$$= 5 - 4$$
$$= 1,$$

revealing that Dodd's salary is \$15K. Palme's method, eq. (4), is much more efficient:

$$\text{SUM}(F \cdot CS \cdot Prof; Sal) = \text{SUM}(F; Sal) - \text{SUM}(F \cdot \overline{CS \cdot Prof}; Sal)$$
$$= \$90\text{K} - \$75\text{K}$$
$$= \$15\text{K}. \quad \blacksquare$$

The foregoing example illustrated individual trackers when the questioner already has identified an individual uniquely. Example 4 shows that the individual tracker may reveal nothing for individuals only partly identified.

Example 4. The questioner knows only that Dodd is a female in the CS Dept. The query system will respond with 2 to the query COUNT($F \cdot CS$), whereupon the questioner knows that "$F \cdot CS$" does not characterize Dodd uniquely. If he tried to guess that Dodd's salary is \$15K, eq. (3) would yield

$$\text{COUNT}(F \cdot CS \cdot \$15\text{K}Sal) = \text{COUNT}(F \cdot \overline{CS} + F \cdot \$15\text{K}Sal)$$
$$- \text{COUNT}(F \cdot \overline{CS})$$
$$= 4 - 3$$
$$= 1.$$

Since this does not reveal which of the two CS females earns \$15K, Dodd's salary has remained secret. ■

5. GENERAL TRACKERS

The individual tracker is based on the concept of using categories known to describe a certain individual to determine other information about that individual. A new individual tracker must be found for each person. The general tracker removes this restriction. It employs a single formula that works for the entire database. No prior knowledge about anyone in the database is required.

A *general tracker* is any characteristic formula T whose query set size is in the restricted subrange $[2k, n - 2k]$ — that is,

$$2k \le \text{COUNT}(T) \le n - 2k. \tag{6}$$

Notice that $q(T)$ is always answerable since its query set size is well within the range $[k, n - k]$. Obviously k must not exceed $n/4$ if a general tracker is to exist at all; in the worst case, $k = n/4$, T is a tracker if and only if $\text{COUNT}(T) = n/2$. By symmetry, T is a tracker if and only if $\bar{T}$ is a tracker. The method of compromise is stated below.

GENERAL TRACKER COMPROMISE. *The value of any unanswerable query $q(C)$ can be computed as follows using any general tracker T. First calculate*

$$Q = q(T) + q(\bar{T}). \tag{7}$$

If COUNT(C) $< k$, *the queries on the right-hand side of this equation are answerable*:

$$q(C) = q(C + T) + q(C + \bar{T}) - Q. \tag{8}$$

Otherwise COUNT(C) $> n - k$ *and the queries on the right-hand side of this equation are answerable*:

$$q(C) = 2Q - q(\bar{C} + T) - q(\bar{C} + \bar{T}). \tag{9}$$

Because at least one of the eqs. (8) *or* (9) *is calculable,* $q(C)$ *can be evaluated with at most* 4 *queries beyond the* 2 *required to find* Q.

PROOF. It is clear that eq. (7) is calculable because T and $\bar{T}$ are both trackers and are answerable. Equations (8) and (9) correspond, respectively, to the cases that $q(C)$ is unanswerable because COUNT(C) $< k$ or COUNT(C) $> n - k$. In proving these equations, we will use the observation that

$$\max[\text{COUNT}(C), \text{COUNT}(T)] \leq \text{COUNT}(C + T) \leq \text{COUNT}(C) + \text{COUNT}(T). \tag{10}$$

Consider the case COUNT(C) $< k$. For this case the definition of tracker (relation (6)) reduces relation (10) to $2k \leq$ COUNT($C + T$) $\leq n - k$. This shows that COUNT($C + T$) is in the range $[k, n - k]$, and hence that $q(C + T)$ is answerable. We may repeat the argument using the tracker $\bar{T}$ and conclude that $q(C + \bar{T})$ is also answerable. Figure 2 uses Venn diagrams to outline a proof of eq. (8). We conclude that COUNT(C) $< k$ implies that eq. (8) may successfully be used to calculate $q(C)$.

In case COUNT(C) $> n - k$, relation (10) shows that $n - k <$ COUNT($C + T$), or that $q(C + T)$ is not answerable and eq. (8) cannot be used. However, by symmetry COUNT($\bar{C}$) $< k$; the previous argument then shows that eq. (8) can be used if C is replaced by $\bar{C}$:

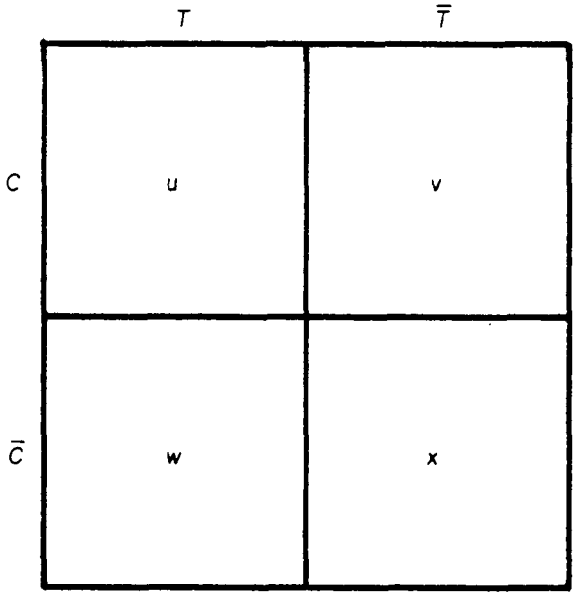
$$q(\bar{C}) = q(\bar{C} + T) + q(\bar{C} + \bar{T}) - Q.$$

By noting that $q(C) = Q - q(\bar{C})$, we can reduce this to eq. (9). ■

The power of the general tracker over the individual tracker should now be clear: Whereas a new individual tracker is required to answer each $q(C)$, a single general tracker suffices to answer every $q(C)$.

Example 5. We will illustrate the general tracker compromise for the database of Table I with $k = 2$. The questioner, who knows that Dodd is a female CS professor, seeks to discover her salary. To be answerable, a query set's size must fall in the range [2, 11], but a general tracker's query set size must fall in the subrange [4, 9]. The formula T = "M" qualifies as a general tracker since COUNT(M) = 7. The questioner applies eq. (7) for counting and summing queries to discover the database size (n) and the total of all salaries (S):

$$\begin{aligned} n &= \text{COUNT}(M) + \text{COUNT}(\bar{M}) \\ &= 7 + 5 \\ &= 12. \end{aligned}$$



$$Q = q(T) + q(\bar{T}) = (u+w) + (v+x)$$
$$= (u+v) + (w+x)$$
$$= q(C) + q(\bar{C})$$

$$q(C+T) + q(C+\bar{T}) = (u+v+w) + (u+v+x)$$
$$= (u+v) + (u+v+w+x)$$
$$= q(C) + Q$$

Fig. 2. Venn diagram showing relations among queries used in the general tracker compromise

$$S = \text{SUM}(M; Sal) + \text{SUM}(\bar{M}; Sal)$$
$$= \$104\text{K} + \$90\text{K}$$
$$= \$194\text{K}.$$

The questioner verifies that Dodd is the only female CS professor by applying eq. (8) with counting queries:

$$\text{COUNT}(F \cdot CS \cdot Prof) = \text{COUNT}(F \cdot CS \cdot Prof + M) + \text{COUNT}(F \cdot CS \cdot Prof + \bar{M}) - n$$
$$= 8 + 5 - 12$$
$$= 1.$$

He then calculates her salary by applying eq. (8) with summing queries:

$$\begin{aligned}\text{SUM}(F \cdot CS \cdot Prof; Sal) &= \text{SUM}(F \cdot CS \cdot Prof + M; Sal) \\ &\quad + \text{SUM}(F \cdot CS \cdot Prof + \bar{M}; Sal) - S \\ &= \$119\text{K} + \$90\text{K} - \$194\text{K} \\ &= \$15\text{K}. \end{aligned}$$ ■

Example 5 illustrated the general tracker used for the same compromise also achieved in Example 3 with the individual tracker. Example 6 illustrates the general tracker when no specific individual is involved.

Example 6. The questioner, who knows that $T =$ "M" is a general tracker for $k = 2$ in Table I, seeks to find the total political contributions paid by persons who are male or are professors. (The answer is \$1185.) First, he applies eq. (7) to find the total of all contributions:

$$\begin{aligned} P &= \text{SUM}(M; Contr) + \text{SUM}(\bar{M}; Contr) \\ &= \$685 + \$510 \\ &= \$1195. \end{aligned}$$

Since $\text{COUNT}(M + Prof) = 11$, the query $\text{SUM}(M + Prof; Contr)$ cannot be answered directly. Since $C + T = (M + Prof) + (M) = C$ for this case, $\text{COUNT}(C + T) = \text{COUNT}(M + Prof) = 11$; therefore queries using $C =$ "$M + Prof$" are not answerable, and the questioner must employ eq. (9):

$$\begin{aligned}\text{SUM}(M + Prof; Contr) &= 2P - \text{SUM}(\overline{M + Prof} + M; Contr) \\ &\quad - \text{SUM}(\overline{M + Prof} + \bar{M}; Contr) \\ &= \$2390 - \$695 - \$510 \\ &= \$1185. \end{aligned}$$ ■

The definition of general tracker T is a sufficient condition for the compromise to work for arbitrary characteristic formulas C. However, it is stronger than necessary. Example 7 illustrates that the compromise may still work for a nontracker T and some (but not all) formulas C.

Example 7. In Table I with $k = 3$, query set sizes must fall in the range [3, 9] to be answerable. The formula $T =$ "$Stat$" is not a general tracker because $\text{COUNT}(Stat) = 3$ is outside the allowable range for trackers [6, 6]. A questioner attempting to apply eqs. (8) or (9) to calculate queries $q(Adm)$ with T as a "tracker" would fail: Equation (8) cannot be applied because $\text{COUNT}(Adm + \overline{Stat}) = 10$, which implies $q(C + \bar{T}) = q(Adm + \overline{Stat})$ is not answerable; eq. (9) cannot be applied either because $\text{COUNT}(\overline{Adm} + Stat) = 11$, which implies $q(\bar{C} + T) = q(\overline{Adm} + Stat)$ is not answerable. On the other hand, both queries

$$\text{COUNT}(F \cdot CS \cdot Prof + Stat) = 4, \qquad \text{COUNT}(F \cdot CS \cdot Prof + \overline{Stat}) = 3$$

are answerable, which implies that eq. (8) can be used to answer questions about Dodd. For example,

$$\begin{aligned}\text{SUM}(F \cdot CS \cdot Prof; Sal) &= \text{SUM}(F \cdot CS \cdot Prof + Stat; Sal) \\ &\quad + \text{SUM}(F \cdot CS \cdot Prof + \overline{Stat}; Sal) - S \\ &= \$75\text{K} + \$134\text{K} - \$194\text{K} \\ &= \$15\text{K}. \end{aligned}$$ ■

The general tracker compromise is clearly a powerful technique. In a later section we will show that almost all databases have a general tracker and we will consider the effort required to find it. We show in Section 6 that two general trackers used together are even more powerful than one.

6. DOUBLE TRACKERS

The general tracker is not guaranteed to work when $k > n/4$, that is, when more than half the range of query set sizes is disallowed. But this does not imply that the database is secure because, corresponding to a given C, there may exist a formula T for which eqs. (7)–(9) work (see Example 7). (There may also exist a decomposition of C for which the individual tracker works.) Even if the database could be proved secure from the general tracker when $k > n/4$, it may be susceptible to compromise by the method of the double tracker.

A *double tracker* is a pair of characteristic formulas (T, U) for which

$$X_T \subseteq X_U, \tag{11a}$$

$$k \leq \text{COUNT}(T) \leq n - 2k, \tag{11b}$$

$$2k \leq \text{COUNT}(U) \leq n - k. \tag{11c}$$

Obviously $k \leq n/3$ if these conditions are to be met at all; in the worst case, $k = n/3$, $\text{COUNT}(T) = n/3$ and $\text{COUNT}(U) = 2n/3$. By symmetry, (T, U) is a double tracker if and only if $(\bar{U}, \bar{T})$ is. The method of compromise is stated below.

DOUBLE TRACKER COMPROMISE. *The value of any unanswerable query $q(C)$ can be computed as follows using any double tracker (T, U). If* $\text{COUNT}(C) < k$, *all queries on the right-hand side of this equation are answerable:*

$$q(C) = q(U) + q(C + T) - q(T) - q(\overline{C \cdot T} \cdot U). \tag{12}$$

Otherwise $\text{COUNT}(C) > n - k$ *and all queries on the right-hand side of this equation are answerable:*

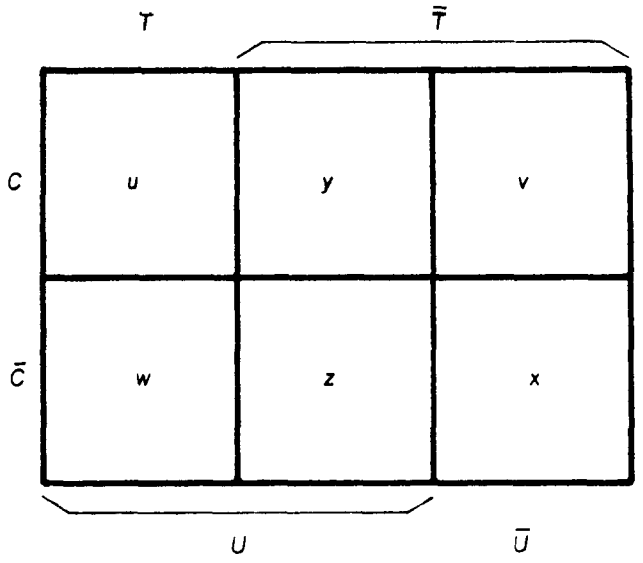
$$q(C) = q(\bar{U}) - q(\bar{C} + T) + q(T) + q(\overline{\bar{C} \cdot T} \cdot U). \tag{13}$$

Because at least one of eqs. (12) *or* (13) *must work, $q(C)$ can be evaluated with at most 7 distinct queries.*

PROOF. The truth of eq. (12) is illustrated by Figure 3. The next two paragraphs explain why the queries on the right-hand side of eq. (12) (or (13)) are answerable.

Consider the case $\text{COUNT}(C) < k$. Using relation (10) with relation (11b),

$$\begin{aligned} k &\leq \max[\text{COUNT}(C), \text{COUNT}(T)] \\ &\leq \text{COUNT}(C + T) \leq k + n - 2k = n - k. \end{aligned} \tag{14}$$



$$
\begin{aligned}
q(C+T)+q(U) &= (u+y+v+w)+(u+y+w+z) \\
&= (u+y+v)+(u+w)+(y+w+z) \\
&= q(C)+q(T)+q(\overline{C\cdot T}\cdot U)
\end{aligned}
$$

Fig. 3. Venn diagram showing relations among queries used in the double tracker compromise

This shows that query $q(C + T)$ is answerable because its query set size is in the range $[k, n - k]$. With the help of Figure 3, we see that

$$\text{COUNT}(\overline{C\cdot T}\cdot U) = \text{COUNT}(U) - \text{COUNT}(C\cdot T). \tag{15}$$

The maximum possible value of this count is COUNT(U) which, by relation (11c), cannot exceed $n - k$. The minimum possible value of this count is min[COUNT(U)] − max[COUNT($C\cdot T$)]; but min[COUNT(U)] = $2k$ by relation (11c) and max[COUNT($C\cdot T$)] ≤ max[COUNT(C)] < k by assumption; hence the minimum of this count is $2k - k = k$. This shows that the query set size for $\overline{C\cdot T}\cdot U$ is in the range $[k, n - k]$, whence $q(\overline{C\cdot T}\cdot U)$ is answerable. Since $q(U)$ and $q(T)$ are answerable by assumption, all the queries on the right-hand side of eq. (12) are answerable.

If COUNT(C) > $n - k$, relation (10) shows that COUNT($C + T$) > $n - k$, whence $q(C + T)$ is unanswerable and eq. (12) will not work. But by symmetry, COUNT($\bar{C}$) < k and the entire previous argument holds with C replaced by $\bar{C}$. In this case Figure 3 shows that

$$q(\bar{C} + T) + q(U) = q(\bar{C}) + q(T) + q(\overline{\bar{C}\cdot T}\cdot U). \tag{16}$$

If we replace $q(\bar{C})$ with $Q - q(C)$ and note that Q can also be expressed as $q(U) + q(\bar{U})$, we can reduce this to eq. (13). ■

Example 8 illustrates the double tracker compromise under a query set size restriction so strong that no general tracker exists.

Example 8. The requirement $k \leq n/4$ precludes a general tracker for Table I when $k = 4$. However, $(T, U) = (Math, Prof)$ is a double tracker in accordance with relations (11):

$$X_{Math} = \{2, 3, 8, 11\} \subseteq \{1, 2, 3, 4, 5, 6, 8, 11\} = X_{Prof};$$
$$\text{COUNT}(Math) = 4 \text{ is in the range } [4, 4]; \quad \text{and}$$
$$\text{COUNT}(Prof) = 8 \text{ is in the range } [8, 8].$$

The questioner may apply eq. (11) for counting queries to verify that Dodd is the only female CS professor:

$$\begin{aligned} \text{COUNT}(F \cdot CS \cdot Prof) &= \text{COUNT}(Prof) && = 8 \\ &\quad + \text{COUNT}(F \cdot CS \cdot Prof + Math) && +5 \\ &\quad - \text{COUNT}(Math) && -4 \\ &\quad - \text{COUNT}(\overline{F \cdot CS \cdot Prof \cdot Math} \cdot Prof) && -8 \\ & && 1 \end{aligned}$$

He may then calculate Dodd's political contribution:

$$\begin{aligned} \text{SUM}(F \cdot CS \cdot Prof; Contr) &= \text{SUM}(Prof; Contr) && = \$1150 \\ &\quad + \text{SUM}(F \cdot CS \cdot Prof + Math; Contr) && +950 \\ &\quad - \text{SUM}(Math; Contr) && -900 \\ &\quad - \text{SUM}(\overline{F \cdot CS \cdot Prof \cdot Math} \cdot Prof; Contr) && -1150 \\ & && \$50 \end{aligned}$$

■

For any general tracker T, relations (11) imply that (T, T) is a double tracker. Indeed, Figure 3 reduces to Figure 2 when $T = U$, and eqs. (12) and (13) reduce to eqs. (8) and (9). If a general tracker exists there is no point in using a double tracker; only when $3k \leq n < 4k$ does the double tracker become interesting.

It is not known whether or not trackers of multiplicity greater than 2 exist for answering the unanswerable when $n/3 < k \leq n/2$. We have not explored this question because the value of the result does not seem justified by its complexity. The single tracker and double tracker define sufficient conditions under which a uniform procedure will calculate any unanswerable $q(C)$. In our experience it is almost always possible, given a C, to find T (and U) so that one of the compromises works even when the database has no general tracker. (This was illustrated in Example 7.) Moreover, a double tracker is ruled out only when less than $\frac{1}{3}$ of the possible query set sizes are observable. Even when k is near $n/2$, individual trackers usually exist [14]. Such severe restrictions on k would ruin the database as a useful source of statistical information without securing the records in it.

7. THE EFFORT TO FIND A TRACKER

There are two questions relating to the security of databases against tracker attacks: How many databases have a tracker? How difficult is finding a tracker? Each question is considered below.

Which Databases Have a Tracker?

Recall that the general tracker is a formula T for which COUNT(T) is in the range $[2k, n - 2k]$. In Appendix 1 we prove:

SUFFICIENT CONDITION FOR GENERAL TRACKER. *Suppose that there are formulas $C_1, \ldots, C_{2k+1}$ whose mutually disjoint query sets collectively exhaust the database. If $n \geq 4k$ there exists a subset I of $\{1, \ldots, 2k + 1\}$ such that the disjunctive formula*

$$T = \sum_{i \in I} C_i \tag{17}$$

is a general tracker.

If some particular category field j contains at least $2k + 1$ distinct values among all the records, then simple formulas like C_i = "category field j has value v_i" can be used to construct a general tracker. Some databases whose records form fewer than $2k + 1$ distinct classes have trackers, others do not (see Appendix 1). In Appendix 1 we also prove:

SUFFICIENT CONDITION FOR DOUBLE TRACKER. *Suppose that there are formulas $C_1, \ldots, C_{2k+1}$ whose mutually disjoint query sets collectively exhaust the database. If $n \geq 3k$ there exists a subset J of a subset K of $\{1, \ldots, 2k + 1\}$ such that the disjunctive formulas*

$$T = \sum_{i \in J} C_i \quad \text{and} \quad U = \sum_{i \in K} C_i \tag{18}$$

form a double tracker (T, U).

These results also imply that the probability that a given database has a general tracker tends to 1 rapidly with n. If we regard the above formulas C_i as defining distinguishable classes of individuals, we see that the probability that the database has a tracker can be less than 1 only if there are fewer than $2k + 1$ classes of individuals. The wider the diversity among the characteristics of individuals, the greater the probability they form at least $2k + 1$ distinct classes. (See Appendix 2.) Such a diversity occurs in practice; for example, Schlörer observed that 98 percent of the records in a medical database were mutually distinguishable by just ten characteristics [14]. Ironically, the utility of the database as a source of statistical information also increases with the diversity among the individuals registered in it.

Because so many databases have general and double trackers, there is little point in studying the probability that an individual tracker can be found.

How Difficult Is Finding a Tracker?

For given C, the time to find out whether an individual tracker exists or not is proportional to the time required to find the required decomposition $C = A \cdot B$. If C is the conjunction of m characteristics of an individual, the search is proportional to the time required to examine each subset of these characteristics, i.e. to 2^m. In a medical database, Schlörer found that 98 of 100 randomly chosen records were uniquely identifiable with 10 or fewer characteristics [14]; in such a case, a questioner with supplementary knowledge of 10 characteristics of an individual could find an individual tracker within $2^{10} = 1024$ queries.

To find a general tracker, the questioner must discover a formula T such that $2k \leq \text{COUNT}(T) \leq n - 2k$. Under the unrealistic assumption that the questioner can inspect all the records in the database, a general tracker can be found in time proportional to at most n^2 (see Appendix 3). Schlörer has recently shown that, if each category-value is equally likely (in its category), then often more than 99 percent of the distinct possible nonempty query sets will correspond to trackers [16]. In other words, a questioner is likely to find a tracker quickly simply by guessing.

Although no definitive study has been made of finding trackers in real databases, these facts suggest that discovering them is not difficult.

8. CONCLUSION

We have studied how to compromise confidential information in statistical databases whose queries use arbitrary characteristic formulas to select subsets of records. Our results apply to a large number of real database systems, including relational ones such as System R or INGRES.

The query system will respond to a query $q(C)$ only if the size of the query set is in the range $[k, n - k]$, where $n \geq 2k$ is the number of records in the database. We considered two kinds of trackers, which are characteristic formulas that help calculate the values of "unanswerable" queries. The individual tracker is a formula $T = A \cdot \overline{B}$ derived from a decomposition of a given formula $C = A \cdot B$, where C identifies a particular individual. The general tracker is any formula T whose query set size is in the range $[2k, n - 2k]$. Whereas a new individual tracker must be found for each new person a questioner desires to investigate, one general tracker can be applied for every person a questioner desires to investigate.

All databases containing $2k + 1$ distinguishable classes of individuals have a general tracker, and many having fewer classes also have trackers. The more diverse the characteristics of individuals, the more interesting is the database as a source of statistical information—and the more likely is the database to have a tracker.

Even if k is large enough to preclude a general or double tracker, the algorithms for compromise may still work for particular choices of the "tracker" formula. Severe restrictions on query set size may seriously impair the utility of the database without securing it.

Even when all access paths to the database are controlled by the query system, the expected time to discover a tracker by trial and error is not high. Trackers are both easy to find and easy to use.

Several avenues are open to building secure statistical databases, not all of which have been explored thoroughly. One possibility is to limit the kinds of characteristic formulas that may be used. This does not appear very promising, since the principle of inclusion and exclusion can be used to calculate responses for arbitrary formulas by adding and subtracting responses using very primitive formulas [14, 17]. A second possibility is to partition the database, arranging that nonempty query sets contain one or more blocks of the partition [21]. Clever querying of such a database can, at best, isolate one of the blocks, but never an individual's record. A third possibility is to give up the requirement that the queries be known functions of specific query sets. For example, data can be

perturbed in unknown ways before being processed by the queries; responses can be perturbed before being reported to the user; query sets can be random samples of the original database [5, 6, 10, 12, 15]. Very large databases are easier to secure than small or medium ones.

The simplicity of these results confirms what has been suspected all along: Compromise is straightforward and cheap. The requirement of complete secrecy of confidential information is not consistent with the requirement of producing exact statistical measures for arbitrary subsets of the population. At least one of these requirements must be relaxed before assurances of security can be believed.

APPENDIX 1. SUFFICIENT CONDITIONS FOR TRACKERS TO EXIST

We will use the following proposition about partitions of integers to prove sufficient conditions for general and double trackers.

PROPOSITION. *Let $y_1, \cdots, y_r$ be $r \geq p + 1$ integers whose sum is $2p$. There exists a subset I of $\{1, 2, \cdots, r\}$ such that $p = \sum_{i \in I} y_i$.*

PROOF. Assume that the indices are chosen so that $1 \leq y_1 \leq \cdots \leq y_r$. Observe that $r \geq p + 1$ implies $y_r \leq p$. If $y_s = m$ is the smallest integer larger than 1, the sum of the integers is at least $s - 1 + m(r - s + 1) = t$. That $t \leq 2p$ implies

$$s \geq (m(r + 1) - 2p - 1)/(m - 1).$$

Using $r \geq p + 1$ and $p \geq y_r \geq m$, we can reduce this to $s \geq m + 1$. In other words, there are at least m 1's.

Now we will show by "downward induction" that the truth of the Proposition for given $r + 1$ implies its truth for $r \geq p + 1$. As a basis, note that $r = 2p$ implies all $y_i = 1$ and that $I = \{1, \ldots, p\}$ is a suitable subset. For $m = y_s$ as defined above, $\{1, y_1, \ldots, y_{s-1}, m - 1, y_{s+1}, \ldots, y_r\}$ has $r + 1$ integers and can, by induction hypothesis, be partitioned into two blocks, in both of which the integers sum to p. We may assume that the integer $m - 1$ is in the same block with a 1 for, if all the 1's are in the other block, we may exchange the integer $m - 1$ with $m - 1$ of the 1's (there are at least m 1's available). We replace the pair of integers $(m - 1, 1)$ with the single integer m. Now both blocks are subsets of the original integers, and in each of them the integers sum to p.

General Tracker

A general tracker is a formula T for which COUNT (T) is in the range $[2k, n - 2k]$, where $n \geq 4k$. Suppose that there are formulas $C_1, \ldots, C_{2k+1}$ whose mutually disjoint query sets collectively exhaust the database; C_i defines the ith "class" of individuals.

Since $n \geq 4k$ we can choose a subset of exactly $4k$ records in which there is at least one record from each class. Denote by y_i the number of records from the ith class. Applying the Proposition with $p = 2k$ and $r = p + 1$, we see that there is a subset I of $\{1, \ldots, 2k + 1\}$ such that exactly $2k$ of the $4k$ records satisfy the disjunctive formula $T = \sum_{i \in I} C_i$. If formula T is applied to the entire database, at most $n - 4k$ additional records can also satisfy T. Therefore,

$$2k \leq \text{COUNT}(T) \leq 2k + (n - 4k) = n - 2k,$$

showing that T is a general tracker.

A simple case in which at least $2k + 1$ classes exist is that some category j contains $r \geq 2k + 1$ distinct values $v_1 < v_2 < \cdots < v_r$ in the database. We can define

$$C_i = \begin{cases} \text{"the value in category } j \text{ is } v_i\text{,"} & 1 \leq i \leq 2k, \\ \text{"the value in category } j \text{ is} > v_i\text{,"} & i = 2k + 1. \end{cases}$$

Classes $1, \ldots, 2k$ correspond to the first $2k$ distinct values in category j, and class $2k + 1$ corresponds to the remaining values.

If there are fewer than $2k + 1$ classes in the database, a tracker may or may not exist. As an example, suppose $k = 2$ and there are four classes with $(y_1, y_2, y_3, y_4) = (1, 1, 1, 5)$; because every characteristic formula matches some subset of these classes, every query set must include no more than 3 or no less than 5 records, whereas a tracker must specify a query set of exactly 4 records in this case. As another example, suppose $k = 2$ and there are two classes with $(y_1, y_2) = (4, 4)$; obviously either formula C_1 or C_2 specifies 4 records and is a tracker in this case.

Double Tracker

A double tracker is a pair of formulas (T, U) such that the query set X_U contains the query set X_T, COUNT(T) is in the range $[k, n - 2k]$, and COUNT(U) is in the range $[2k, n - k]$. We suppose that there are formulas $C_1, \ldots, C_{2k+1}$ whose mutually disjoint query sets exhaust the database, and that $3k \leq n < 4k$ so that no general tracker can exist.

Choose a subset of exactly $3k$ records in which there is at least one record from each class. Let y_i denote the number of records in the ith class. At least $k + 2$ of these y_i must be 1, which implies that a subset of $2k$ records come from $k + 1$ classes. Let K be the indices of these classes. Applying the Proposition with $p = k$, there must be a subset J of K whose classes include k records. Let T and U be the disjunctive formulas

$$T = \sum_{i \in J} C_i \quad \text{and} \quad U = \sum_{i \in K} C_i.$$

Note that

$$k = \sum_{i \in J} y_i \quad \text{and} \quad 2k = \sum_{i \in K} y_i.$$

Since J is contained in K, the query set X_T is contained in X_U. If T is applied to the entire database, at most $n - 3k$ additional records can also satisfy T; thus

$$k \leq \text{COUNT}(T) \leq k + (n - 3k) = n - 2k.$$

Similarly, if U is applied to the entire database, at most $n - 3k$ records additional can also satisfy U; thus

$$2k \leq \text{COUNT}(U) \leq 2k + (n - 3k) = n - k.$$

We conclude that (T, U) is a double tracker.

APPENDIX 2. PROBABILITY THAT THE DATABASE CONTAINS A TRACKER

A class is a set of records with identical category fields. With n records there are at most n nonempty classes. If we suppose that each individual is independent and equally likely to belong to the nonempty classes, we can estimate the probability that the database has a tracker. Let S be a subset of $\{1, \ldots, n\}$ containing $2k$ of the nonempty class indices. Let $z_i = 1$ if individual i is a member of any class of S, and $z_i = 0$ otherwise. Note that $\Pr[z_i = 1] = 2k/n$, which implies that the mean and variance of z_i are

$$\bar{z}_i = 2k/n, \qquad \sigma_i^2 = (2k/n)(1 - 2k/n).$$

Define $Z = z_1 + \cdots + z_n$; Z is the number of individuals in the classes of S. Its mean and variance are

$$\bar{Z} = 2k, \qquad \sigma^2 = 2k(1 - 2k/n).$$

Since Z is the sum of independent random variables, it is approximately normal with the cumulative distribution

$$\Pr[Z \leq N] = \Phi((N - \bar{Z})/\sigma),$$

where $\Phi(u)$ is the cumulative distribution of a normal random variable with mean 0 and variance 1. The probability that all n records fall in the $2k$ classes of S is an upper bound on the probability that no tracker exists,

$$\Pr[\text{no tracker}] < \Pr[Z = n] \cong \Phi((n - \bar{Z})/\sigma) - \Phi((n - 1 - \bar{Z})/\sigma).$$

This quantity has been determined for a particular subset S; however, the symmetry of the problem implies that $\Pr[Z = n]$ is the same for every S and is, therefore, an unconditional probability. For $n = 4k$, the worst case, this reduces to

$$\Pr[Z = n] \cong \Phi(n^{1/2}) - \Phi(n^{1/2} - 2/n^{1/2}),$$

which approaches 0 very rapidly for increasing n. Even for $n = 9$, this expression is less than 0.01.

APPENDIX 3. ALGORITHM TO FIND A GENERAL TRACKER

Assume that there n records and a fixed number of category fields. In $O(n^2)$ time one can sort the records by category fields and count the size of each distinguishable group of records. Let $y_1, y_2, \ldots, y_r$ ($r \leq n$) denote these counts and $C_1, C_2, \ldots, C_r$ be corresponding formulas. Note that the y_i sum to n and that every formula's query set can also be specified by a subset of the C_i.

In $O(n^2)$ time one can construct a Boolean matrix $B[i, j]$ for $1 \leq i \leq r$ and $0 \leq j \leq n$, such that $B[i, j]$ denotes the proposition "there exists a subset of $y_1, \ldots, y_i$ whose sum is j." As a basis, $B[1, j] = 1$ if and only if $j = 0$ or $j = y_1$. Inductively, $B[i + 1, j] = 1$ if $B[i, j] = 1$, or if $y_{i+1} \leq j$ and $B[i, j - y_{i+1}] = 1$. A general tracker exists if and only if there is a 1 in some column j, $2k \leq j \leq n - 2k$; note from Appendix 1 that $r \geq 2k + 1$ implies that a general tracker exists. (This is a minor adaptation of an algorithm given by Garey and Johnson for partitioning a set of integers [7].)

Assuming that $B[i, j] = 1$ for some j, $2k \leq j \leq n - 2k$, we can in $O(n)$ time construct a set S of indices such that

$$\sum_{i \in S} y_i = j,$$

whence the disjunctive formula

$$\sum_{i \in S} C_i = T$$

is a general tracker. To construct S, repeat this step until $j = 0$: Reduce i until $i = 1$ or $B[i - 1, j] = 0$; then add i to S, set j to $j - y_i$, and set i to $i - 1$.

The total time to accomplish all these tasks is $O(n^2)$.

ACKNOWLEDGMENTS

We are grateful to Douglas Comer and Michael O'Donnell for insights into the proof of the Proposition in Appendix 1, and to Jan Schlörer for encouragement and for noting an error in an earlier draft. We are also grateful to D. S. Johnson for pointing out the $O(n^2)$ algorithm for finding a general tracker. Finally, we are grateful to the referees for their comments and suggestions.

REFERENCES

1. ASTRAHAN, M.M., ET AL. System R: Relational approach to database management. *ACM Trans. Database Syst. 1,* 2 (June 1976), 97-137.
2. CHAMBERLIN, D.D., AND BOYCE, R. SEQUEL: A structured English query language. Proc. ACM SIGMOD Workshop on Data Description, Access, and Control, May 1974, pp. 249-264.
3. CHIN, F.Y. Security in statistical data bases for queries with small counts. *ACM Trans. Database Syst. 3,* 1 (March 1978), 92-104.
4. DOBKIN, D., JONES, A.K., AND LIPTON, R.J. Secure databases: Protection against user inference. Res. Rep. No. 65, Dept. Comptr. Sci., Yale U., New Haven, Conn., April 1976. To appear in *ACM Trans. Database Syst.*
5. FELLEGI, I.P. On the question of statistical confidentiality. *J. Amer. Statist. Assoc. 67,* 337 (March 1972), 7-18.
6. FELLEGI, I.P., AND PHILLIPS, J. L. Statistical confidentiality: Some theory and applications to data dissemination. *Annals Econ. Soc'l Measurement 3,* 2 (April 1974), 399-409.
7. GAREY, M.R., AND JOHNSON, D. S. Strong NP-completeness results: Motivation, examples, and implications. *J. ACM 25,* 3(July 1978), 499-508.
8. HANSEN, M.H. Insuring confidentiality of individual records in data storage and retrieval for statistical purposes. Proc. AFIPS 1971 FJCC, Vol. 39, AFIPS Press, Montvale, N.J., pp. 579-585.
9. HAQ, M.I. Security in a statistical data base. *Proc. Amer. Soc. Inform. Sci. 11* (1974), 33-39.
10. HOFFMAN, L.J., AND MILLER, W.F. Getting a personal dossier from a statistical data bank. *Datamation 16,* 5 (May 1970), 74-75.
11. KAM, J.B., AND ULLMAN, J.D. A model of statistical databases and their security. *ACM Trans. Database Syst. 2,* 1 (March 1977), 1-10.
12. NARGUNDKAR, M.S., AND SAVELAND, W. Random rounding to prevent statistical disclosure. *Proc. Amer. Statist. Assoc., Soc. Statistics Sect.* (1972), 382-385.
13. PALME, J. Software security. *Datamation 20,* 1 (Jan. 1974), 51-55.
14. SCHLÖRER, J. Identification and retrieval of personal records from a statistical data bank. *Methods of Inform. in Medicine 14,* 1 (Jan. 1975), 7-13.
15. SCHLÖRER, J. Confidentiality of statistical records: A threat monitoring scheme for on-line dialogue. *Methods of Inform. in Medicine 15,* 1 (Jan. 1976), 36-42.
16. SCHLÖRER, J. Union tracker and open statistical databases. Rep. TB-IMSD 1/78, Institut für Medizinische Statistik und Dokumentation, Universität Giessen, June 1978.

17. SCHWARTZ, M.D. Inference from statistical data bases. Ph.D. Th., Dept. Comptr. Sci., Purdue U., W. Lafayette, Ind., Aug. 1977.
18. SCHWARTZ, M.D., DENNING, D.E., AND DENNING, P.J. Linear queries in statistical data bases. TR-216, Dept. Comptr. Sci., Purdue, U., W. Lafayette, Ind., Nov. 1976. To appear in *ACM Trans. Database Syst.*
19. SCHWARTZ, M.D., DENNING, D.E., AND DENNING, P.J. Securing data bases under linear queries. Information Processing 77, North-Holland Pub. Co., Amsterdam, 1977, pp. 395–398.
20. STONEBRAKER, M., WONG, E., KREPS, P., AND HELD, G. The design and implementation of INGRES. *ACM Trans. Database Syst. 1,* 3 (Sept. 1976), 189–222.
21. YU, C.T., AND CHIN, F.Y. A study on the protection of statistical data bases. ACM SIGMOD Conf. Manage. of Data, Toronto, Canada, Aug. 1977, pp. 169–181.
22. WEIDE, B. A survey of analysis techniques for discrete algorithms. *Comptng. Surveys* 9, 4 (Dec. 1977), 291–313.

Received November 1977; revised November 1978