DISCUSSION

David L. Zalkind, Department of Health, Education and Welfare

The papers by Sande and by Dalenius and Reiss deal with two different aspects of database confidentiality. The Sande paper is similar in spirit to the paper by Cox presented earlier in this session. The basic question that is the focus of consideration is whether a given pattern of suppression of table entries is satisfactory to preserve confidentiality according to some predetermined criterion. One suggested criterion is that for each suppressed entry there is a sufficiently large range of values such that no inferrences about the true value can be made insided this range. The problem may stated as follows:

Suppose the data table of interest is m by n, exclusive of the marginals. Let I be the set of mn entry positions. Let P be the set of positions that are published, so $P \subset I$. Let S be the set of entry positions that correspond to "sensitive" data that must be protected (i.e., can only be inferred to at best some specified range), so $S \subset I - P$. Let the set of positions corresponding to non-sensitive data that must be suppressed in order to protect sensitive data, be represented by N, so $N = I - P - S$.

As explained by both Sande and Cox, the original table can be represented in a special form called a transportation problem:

$$T X = B , \qquad (1)$$

where T is a particular unimodular matrix, X represents the table entries, and B is an $m + n$ column vector consisting of the m row marginals and n column marginals. (See Sande or Cox for details.)

If one partitions the data entries in the vector X to correspond to the three sets P, N, and S, and similarly rearranges columns of T, then equation (1) can be rewritten as

$$T_P A_P + T_N X_N + T_S X_S = B \qquad (2)$$

where A_P are the true (published) values of the data. Subtracting $T_P A_P$ from both sides of equation (2), we still have a transportation problem, since $B - T_P A_P$ is a constant vector.

Now, sensitive data is considered to be protected if and only if for every $k \in S$, there exist feasible solutions to each of the individual inequalities

$$X_k \leq L_k \qquad (3)$$

and

$$X_k \geq U_k , \qquad (4)$$

where L_k and U_k are the "protective" bounds. If $\#(S)^k$ is the number of elements in S, then there are $2\ \#(S)$ "capacitated transportation problems" created by individually appending each of the constraints in inequalities (3) and (4) to the transportaition problem (2).

For a given set S and a given set N, the remaining problem is to determine as efficiently as possible if each of the capacitated transportation problems has a feasible solution. Fortunately, it is quite easy to find such a solution to a transportation problem, and not much more difficult to find one for a problem with a single added capacity constraint.

One contrast between the Sande and Cox papers is that Sande explicitly considers the combinatorial possibilities that might lead to inadvertant disclosures of sensitive data through algebraic manipulation of the inequalities implicit in the transportation problem whereas Cox treats the transportation problem as a "black box" that will let him know if too much information has been revealed. This discussant believes that Cox's approach will prove to be more efficient, although both have merit in yielding insight about the underlying problem.

Perhaps Cox's approach may be made even more efficient by invoking additional properties available in standard approaches to solving transportation problems. For example, if one were to create special artificial sets of costs corresponding to entries in the transportation problem, the initial feasible solution that is found by standard mathematical programming methods will simultaneously satisfy many of the capacity constraints (3) and (4) and thus greatly reduce the number of feasible solutions that need to be found. Thus, by using an objective function with large costs corresponding to sensitive variables, one may find an initial feasible solution satisfying most of the constraints (3). Furthermore, by perturbing this artificial objective function, such as reducing costs corresponding to variables for which constraints (3) have already been satisfied and increasing costs for those which have not, one may be able to find a new feasible solution satisfying most of the remainder of constraints (3).

Similarly, one could create an artificial objective function with large costs attributed to variables corresponding to nonsensitive suppressed variables and thus satisfy several constraints (4) at one time. One would then change this objective function in a manner similar to that above in order to obtain other feasible solutions satisfying other constraints (4).

Note that we have only used objective functions as an aid in finding useful feasible solutions, but have not found "optimal" solutions. An unresolved problem is the construction of "maximal" sets P of published data. Perhaps one could develop a hybrid algorithm that whould recognize when a variable can be transferred from the set N of nonsensitive suppressed data to the set P. (This discussant feels that a simple "myopic" algortithm can be developed to find good solutions, but that an integer program may be required to

find optimal solutions. Unfortunately, integer programs may require a considerable amount of computation to solve.)

While the papers by Cox and by Sande focus on testing data suppression patterns, and by implication point up the necessity of finding an efficient approach to constructing "good" suppression patterns, the paper by Dalenius and Reiss protects table entries in a different manner. Rather than suppress entries, Dalenius and Reiss propose that entries be rearranged in such a way that marginals up to a certain order of cross-tabulation are unaffected, while values for individual respondents cannot be inferred with certainty since there is a positive probability of "sufficient" magnitude that any given entry in the published table is not the same as the entry in the original table. This approach is applicable to data from individual respondents with relatively few categorical responses for each data item, rather than to the presumably aggregated data considered in Cox and in Sande. A major advantage of the data swapping approach is that unaggregated data my be released (or inadvertently revealed) without compromising individuals responses or respondents. Thus, the notion of data swapping is quite intriguing.

The crux of the Dalenius and Reiss paper is an argument showing that such a data swap exists with any desired probability (less than 1) if there are a sufficient number of respondents in the survey.

In developing a non-constructive argument of the kind in the paper, one must be careful that simplifying, bounding, and approximating assumptions actually yield results that are as strong as one intends. For example, simplifying approximations should not be made without checking that needed inequalities are preserved.

Assuming the validity of the derivations presented and anticipating the tightening of some of the bounds, one looks forward to the presentation of some computational work to demonstrate the feasibility of the approach with actual data sets. However, there is some concern that a valid data switching algorithm may involve so much combinatorics that it is equivalent to an integer program and thus take considerable computational time even for a moderate size data set.