DISCLOSURE ANALYSIS AND CELL SUPPRESSION

Lawrence H. Cox, U. S. Census Bureau

The various approaches to the problem of statistical confidentiality include rolling-up, random rounding, randomized response and cell suppression. In certain demographic censuses and in large-scale economic censuses and surveys in which there are many levels of aggregation detail over which the data is inhomogeneously distributed, techniques of cell suppression are generally employed. Historically, this process was accomplished by subject-matter analysts virtually by hand. The enormity of the problem and increased technology have brought the problem into the realm of large-scale data processing. This paper is a thumbnail sketch of a disclosure analysis and cell suppression strategy which is compatible with the processing and tabulation runstream and which maintains confidentiality while minimizing over-suppression. This research is on-going and owes much of its present form to previous work of James P. Corbett.

All statistics are considered as definable in terms of two independent schemes of aggregation, those of geography and subject matter. Each level of geography and subject matter analysis is assigned a unique code. Each publication statistic is therefore the aggregate of data elements over a unique sub-file of the complete data file, namely that generated by sorting the file on the corresponding geographic and subject matter codes. From this perspective, the processing of the data may be viewed as a series of sorts and aggregations performed on sub-files of the data file, so that the process is realizable conceptually in the form of the organizing lattice of Figure 1. This lattice is the graphical representation of the Boolean algebra generated by the geo- and subject-matter codes.

Subject matter considerations define certain statistics as sensitive. In the U.S. Census of Manufactures, for example, publication statistics are the aggregates over m companies of certain quantative attributes of those companies. A statistic S is defined as sensitive if the aggregate of the corresponding attribute over n of those m companies is equal to or greater than k% of the total value of S, for fixed values of n and k. This is referred to as the "n-company, k% rule" and the values of n and k are themselves treated as U. S. Census Bureau-confidential information as an additional safeguard to privacy. A sensitive statistic S is of course suppressed from publication, but its value can still be estimated by complementation -- i.e., by subtracting the sum of all published statistics appearing in the same row as S from the marginal total of that row. As close estimation of a sensitive statistics is no less a breach of confidentiality than the precise calculation of its value, two bounds, L and U, are computed for each sensitive statistic S. A state of disclosure is said to exist if, on the basis of the publication tables, it is possible to compute any estimate E of the value of S for which $L<E<U$. If a state of disclosure exists within a row, additional suppressions must be made within that row to render the row disclosure-free. These suppressions have themselves become sensitive statistics of a sort, for estimation of their values allows for estimation of the sensitive statistics they were suppressed to protect. Thus, additional suppressions may be required in the other rows in which these protecting statistics appear, and so on. As each publication statistic generally appears in two or three rows and as a row may contain no, one or several suppressed statistics, the task of rendering a set of publication tables disclosure-free emerges as a complex geometrical and logical problem, hence one best approached mathematically.

Each point in the lattice of Figure 1 represents a collection of publication tables at the indicated level of aggregation, together with their associated row marginal totals. The processing of the data and the disclosure analysis of the resulting set of publication tables is then represented as one enormous top-to-bottom pass through the lattice. At any stage of this process (lattice point), the tables are constructed and the individual cells are analyzed to identify the sensitive statistics. These, together with the identification of those row marginal totals which, as cells, were suppressed at one higher level of aggregation, provide the information necessary to identify the disclosed rows. An appropriate cell suppression pattern involving the minimum number of additional suppressions possible is chosen to render the publication table disclosure-free. The enumeration and construction of all such patterns follows from the

Theorem: In a given publication table, assume that one additional suppression in a row suffices to protect any sensitive statistic in that row. Let R denote the number of disclosed horizontal rows and let C denote the number of disclosed vertical rows in the table. If $R = C = 1$, then the table can be rendered disclosure-free by means of at most 3 additional suppressions. Otherwise, $M = \text{Max}\{R,C\}$ additional suppressions are necessary and sufficient. Moreover, the number of distinct such suppression schemes is almost factorial in M.

As suppressions are made only within the table and not along the marginal rows, this process proceeds linearly through the lattice and hence never loops back on itself, thereby minimizing the possibility of "over-suppression" and guaranteeing the efficiency of the process, both in terms of cost and output.

The major advantage of this approach is its generality -- the subject matter analysis by which the sensitive statistics are defined is completely independent from the geometric techniques employed to render the tables disclosure-free. Thus, this approach lends itself to the processing and disclosure analysis of any statistical data for which the use of cell suppression techniques is deemed appropriate. This includes all economic censuses and, in principle, applies to the disclosure problem of reconciling special

tabulations with U. S. Census publications. The methodological gap to be bridged in this case is the discovery of a theoretically sound, optimal strategy for cell suppression in several dimensions. This is currently under investigation. This procedure generally yields the best-possible upper and lower bounds computable for any suppressed cell and gives researchers a vantage point from which to study and evaluate the impact of different methods of geographic and subject matter aggregation of data on the information content of the resulting publications.

Figure 1

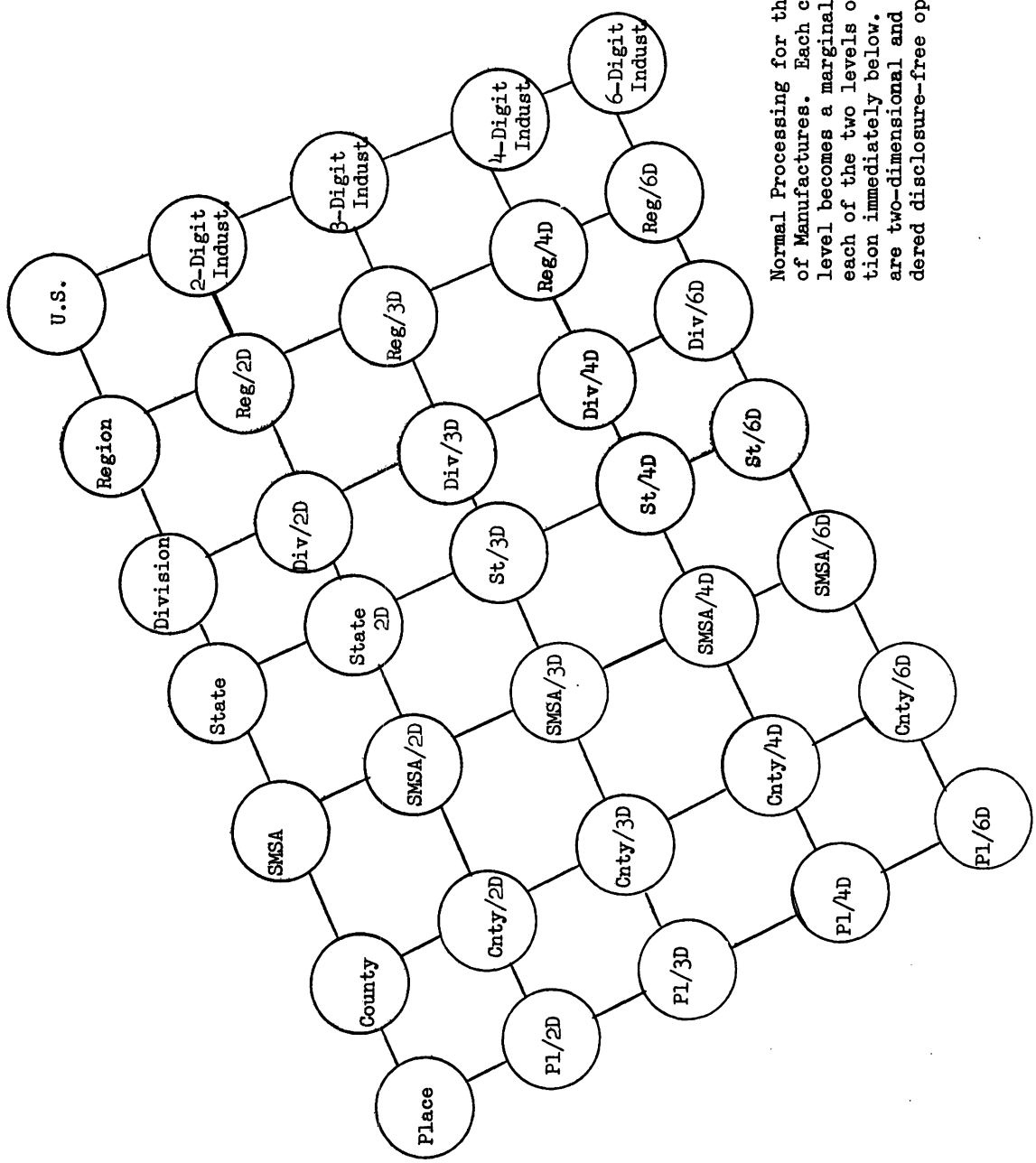
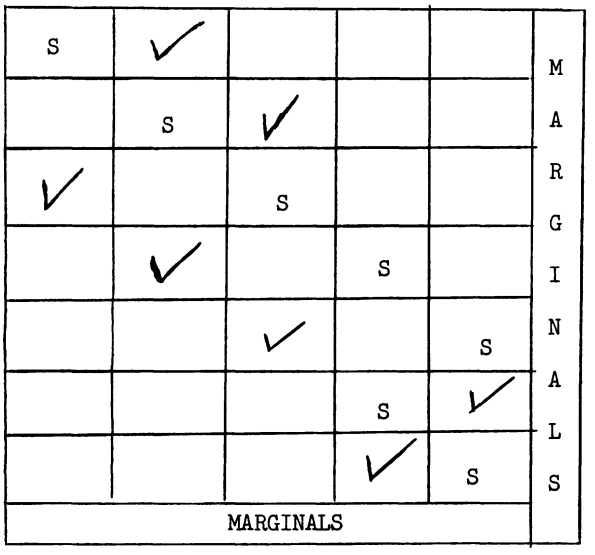


Figure 2



Each "S" represents a sensitive statistic in the above table of statistics. A minimum of seven additional suppressions is required to render this table disclosure-free and, assuming that one additional suppression in each disclosed row or column will render that row disclosure-free, there are as many as 8,492 different suppression schemes that will do so. The checkmarks are an example of one such scheme.