Ronald M. Dopkowski. Demographic Surveys Division. U.S. Bureau of the Census, March 27.

LUNT, T. F., DENNING, D., SCHELL, R. R., HECKMAN, M. and SHOCKLEY, W. R. (1988). Element-level classification with A1 assurance. *Computers and Security* **7** 73–81.

LUXEMBOURG INCOME STUDY NEWSLETTER (1989). Timothy M. Smeeding, Project Director. July, Vanderbilt Univ., Nashville, Tenn.

MCGUCKIN, R. and NGUYEN, S. (1988a). Use of 'surrogate files' to conduct economic studies with longitudinal microdata. In *Proceedings of the Third Annual Research Conference*. Bureau of the Census, Washington, D.C.

MCGUCKIN, R. and NGUYEN, S. (1988b). Public use microdata: disclosure and usefulness. Center for Economic Studies Discussion Paper. CES 88-3, September. U.S. Census Bureau, Washington, D.C.

OFFICE OF FEDERAL STATISTICAL POLICY AND STANDARDS (1978). Report on Statistical Disclosure and Disclosure-Avoidance Techniques. Statistical Policy Working Paper 2, U.S. Department of Commerce. U.S. Government Printing Office, Washington, D.C.

PAASS, G. (1988). Disclosure risk and disclosure avoidance for microdata. *Journal of Business and Economic Statistics* **6** 487–500.

PALLEY, M. A. and SIMONOFF, J. S. (1986). Regression methodology based disclosure of a statistical database. In *Proceedings of the Section on Survey Research Methods* 382–387. Amer. Statist. Assoc., Alexandria, Va.

PEARSON, R. W. (1987). Researchers' access to U.S. federal statistics. *Items* **41** 6–11.

RAINWATER, L. and SMEEDING, T. M. (1988). The Luxembourg Income Study: the use of international telecommunications in comparative social research. *ANNALS. AAPSS* **495** 95–105.

ROBERTS, H. V. (1986). Comment on "Disclosure-limited data dissemimation" by G. T. Duncan and D. Lambert. *J. Amer. Statist. Assoc.* **81** 25–27.

SHOSANI, A. (1982). Statistical databases: characteristics, problems, and some solutions. In *LBL Perspective on Statistical Database Management* 3–28. Lawrence Berkeley Laboratory, Univ. California, Berkeley.

SPRUILL, N. L. (1983). The confidentiality and analytic usefulness of masked business microdata. In *Proceedings of the Section on Survey Research Methods* 602–607. Amer. Statist. Assoc., Alexandria, Va.

STRUDLER, M., OH, H. L. and SCHEUREN, F. (1986). Protection of taxpayer confidentiality with respect to the tax model. In *Proceedings of the Section on Survey Research Methods*, 375–381. Amer. Statist. Assoc., Alexandria, Va.

SUBCOMMITTEE ON DISCLOSURE AVOIDANCE TECHNIQUES (Federal Committee on Statistical Methodology) (1978). Statistical Working Paper 2, Federal Statistical Policy and Standards, U.S. Department of Commerce. Government Printing Office, Washington, D.C.

SULLIVAN, G. and FULLER, W. A. (1989). The use of measurement error to avoid disclosure. Presented at the Annual Meeting of the American Statistical Association.

WARNER, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *J. Amer. Statist. Assoc.* **60** 63–69.

WOLF, M. K. (1988). Microaggregation and disclosure avoidance for economic establishment data. In *Proceedings of the Business and Economics Statistics Section* 355–360. Amer. Statist. Assoc., Alexandria, Va.

Comment

Lawrence H. Cox

This article has many ideas to offer, and I am mostly in agreement with the authors' scenario for the future. I will limit my comments to expanding upon one technical area and suggesting a policy area not discussed by the authors.

MATRIX MASKS

I applaud the characterization of certain data masking techniques in terms of matrix operations $AXB + C$ on the original data matrix X, where (A, B, C) may depend on X. This characterization offers brevity in expresion and the opportunity to study and compare matrix masking methods using standard tools. It will facilitate the development, analysis and maintenance of computer programs to perform data masking, and it also may attract the attention of a wider class of researchers to problems in data masking.

However, the authors observe that the following are not representable as matrix masks of the form $AXB + C$: attribute-specific aggregation over (selected sets of) records; data swapping among some, but not all, attribute fields; (randomly) rounding (all) entries of X; multiplication by random noise generated independently; data grouping; and truncation. These data masks indeed can be represented as matrix masks, in some cases by generalizing the definition of matrix mask to include sums or repeated application of elementary matrix masks $M = AXB + C$ and in other cases by allowing more general arithmetic. Assume henceforth that X is an $m \times n$ matrix.

Lawrence H. Cox is former Director, Board on Mathematical Sciences, Commission on Physical Sciences, Mathematics and Applications, National Academy of Sciences, 2101 Constitution Avenue, Washington, D.C. 20418.

First, (random) rounding. Data rounding is a somewhat restricted form of data perturbation. (Random) data perturbation (i.e., adding (random) noise) can be represented as a matrix mask $M = X + C$, where C contains the perturbation values (equal to zero for all entries not subject to perturbation). The same can be done for (random) data rounding: After determining (randomly) whether an entry is to be rounded down or up, set the corresponding entry of C equal to the difference between the rounded and original values.

For the case of multiplication by independently distributed random values, simply replace X by $\log X$, thereby reducing the problem to that of random additive data perturbation. To avoid confusion when applying other masks to X, one could perform this masking operation last.

In what follows, we adopt the following notation: I denotes the matrix with ones along its main diagonal and zeroes elsewhere; J denotes the matrix of all ones, and Z the matrix of all zeroes; U_{ij} is the matrix with a one in entry (i, j) and zeroes elsewhere. We will specify the dimensions of these matrices as they are used in each case.

For grouping, first represent the categorical data as follows. Assign to each of the c categories to be grouped a column of X; for each record (row) of X, place a 1 in the column corresponding to the correct category for this respondent (or a zero if none of these categories applies), and place a zero in the other $(c - 1)$ columns. Assume that these are the first c columns of X. Grouping can be represented forming $M = XB$, where B is an $n \times (n + 1 - c)$ matrix defined as follows: The first column of B contains ones in the first c rows and zeroes in the remaining $(n - c)$ rows; the remaining $(n - c)$ columns of B consist of a $c \times (n - c)$ Z-matrix in the first c rows, and an $(n - c) \times (n - c)$ I-matrix in the last $(n - c)$ rows. The grouped category appears as the first column of M, the original categories having been deleted.

Truncation (to the value t) for, let's assume, nonnegative integer values v could be performed by using two columns of X, containing nonnegative integer entries f and r, with $r < t$, corresponding to $v = ft + r$. The final value for each record could be achieved when constructing the $m \times (n - 1)$ matrix M from X by using the method of the preceding paragraph to delete one of these X-columns and placing $(\mathrm{Max}\{r, -1 + (t + 1)^v\})\mathrm{mod}(t + 1)$ in the other. This, and the use of logs above, involves using more than standard arithmetic operations, but that really is the point: If matrix masking is to prove a useful representation for data masking, then it should be defined in the broadest possible way. In particular, as demonstrated below, matrix masks could usefully be defined as sums and iterates of what I refer to as the *elementary matrix masks* $M = AXB + C$ defined by the authors.

The underlying problem in the troublesome cases presented by the authors is how to perform masking operations only on some records and some attributes, while leaving other records and attributes fixed: say, for simplicity, on the first p rows and the first q columns. To do so, we use matrix operations to create an $m \times n$ matrix X' whose uppermost left $p \times q$ block is identical to that of X but that contains zeroes elsewhere. The matrix X' is given by $X' = AXB$, as follows: A is an $m \times m$ matrix consisting of a $p \times p$ I-matrix in the uppermost left block and zeroes elsewhere; similarly, B is an $n \times n$ matrix consisting of a $q \times q$ I-matrix in its uppermost left block and zeroes elsewhere.

To perform attribute-specific aggregation (for the first q attributes, aggregated over the first p records), construct the matrix $M' = JX'$ whose entries are the sums by attribute of the entries of X'. If averages are desired, then use $(1/p)J$ instead. Then, $M = M' + X - X'$ is the desired matrix.

To perform data swapping among, say, the first two attribute fields for the first p records, form X' as above $(q = 2)$. Let $M' = X'B$, where B is the $n \times n$ matrix $B = U_{12} + U_{21}$. Again, $M = M' + X - X'$ is the desired matrix.

One may represent combinations of data masks, applied selectively to rows and columns—such as rounding for some attributes and rows and suppressing others—by repeated application of elementary matrix masks (A, B, C) to (masked) matrices $M = M' + X - X'$.

My point above is not to develop a calculus for matrix masking—although that would be useful—but rather is to reinforce the authors' very good idea of representing data masks in unified, familiar way. Indeed, there has been work in graph theory and matroid theory applied to aggregates represented as matrices, thereby eliminating the need for rank computations; and, if suppressed or collapsed data were replaced by interval data (representing confidentiality protected "best" estimates of suppressed cells), then perhaps methods of interval arithmetic could be applied to the analysis of masked matricies.

GAINING PUBLIC TRUST AND SUPPORT

An issue not discussed by the authors is the development of public trust in the statistical system. I do not believe that we will be successful in relaxing unrealistic or unenforceable confidentiality statutes—thereby relieving some of the tension between privacy and access concerns, while maintaining high levels of participation and truthfulness in surveys—without public support.

I believe the time has arrived for the statistical system to "go public" with confidentiality and data access issues. This is based on my opinion that both the policy debate and the development of technical solutions to disclosure protection problems within the statistical system have matured sufficiently to be analyzed and discussed at a general level. This was not the case 10 years ago. There are dangers to raising these issues, however, that must be kept in mind: By raising the issue, it can be made salient in a way that frightens the public; also, we run the risk of confusing people with arcane, inconclusive or contradictory technical and legal information (thereby eroding their confidence in a different way).

The approach should be to communicate the importance of reliable and varied statistics to the society and the economy, while instilling confidence that individual respondents have rights, the protection of which is the bedrock of the statistical system. The approach should first be made through influence groups: advocates for privacy; groups representing the disadvantaged (e.g., the homeless) or those at risk (e.g., AIDS); those concerned with the rights of individuals (e.g., ACLU); the press; and those concerned with the political process. Avenues to many of these groups exist already within the normal workings of the statistical system. Well- and nontechnically articulated arguments need to be developed and discussed with these groups, leading perhaps to experiments of one kind or another, and, ultimately, consensus and change. If these groups are convinced and, to a degree, become advocates for the statistical system on issues of privacy and data access, I believe the support of the public at large will follow.

Within these deliberations, it is important to maintain a focus on legislative issues. Laws do not prescribe how statistical agencies are to design questionnaires and samples, estimate parameters or edit questionnaires and impute missing or faulty data; yet regarding confidentiality, many laws are absolute, one-side in assigning penalties and, although written in the absence of technical information, exert a driving influence on agencies' confidentiality practices. Most agree that responsibility for disclosure protection should, like the data, be shared between the data provider and the data user. This strikes me as an issue easily understood and potentially supportable by outside groups.

The Duncan–Pearson article does a good job of presenting the mounting issues faced by the statistical system along the data confidentiality/data access front. It is readable outside the statistical community, and that is important if we are to broaden the discussion, as I suggest be done.

Comment

Sallie Keller-McNulty

I would like to commend Duncan and Pearson for their contribution on the very important topic of data access and confidentiality. I am pleased that *Statistical Science* has had the foresight to publish such an article, and I hope that many researchers will read and react to the material. I have no disagreements with the opinions expressed in this manuscript, but I would like to bring more attention to a few points that were made.

First, I would like to comment on the various ways that disclosure has been conceptualized. In particular, attention has been focused on inferring *attribute* values. I contend that, with the database systems as a data storage and access medium, we need to also be concerned with the direct disclosure of *relationships* between attributes. We have been conditioned to view data as a file or rectangular array where the columns represent attributes, the rows represent data records (one for each respondent), and the entries within a row represent attribute values. Attribute disclosure is conceptualized as inferring an element of this array. In this setup, hypotheses about the relationships between attributes are validated through analysis of the data. In a database, relationships among attributes are contained in the schema, or logical structure, of the system. Relationships as well as attribute values are considered objects (i.e., encapsulation of values with their semantic meanings). Disclosure in a database system can be defined as inferring an

Sallie Keller-McNulty is Associate Professor of Statistics, Kansas State University, Manhattan, Kansas 66506.