*Statistical Science*
1991, Vol. 6, No. 3, 219–239

# Enhancing Access to Microdata while Protecting Confidentiality: Prospects for the Future

**George T. Duncan and Robert W. Pearson**

*Abstract.* This article presents a scenario for the future of research access to federally collected microdata. Many researchers find access to government databases increasingly desirable. The databases themselves are more comprehensive, of better quality and—with improved database management techniques—better structured. Advances in computer communications enable remote access to these databases. Substantial gains in the performance/cost ratio of computers permit more sophisticated analyses—including ones based on statistical graphics, identification of extreme or influential values, record linkage and Bayesian regression methods.

At the same time, the individuals and institutions that provide the data residing on government databases—as well as the agencies who sponsor the collection of such information—are becoming increasingly aware that the same technologies that extend analytical capabilities also furnish tools that threaten the confidentiality of data records.

As the broker between the data provider and the data user, government agencies are under increased pressure to implement policies that both increase data access and ensure confidentiality. In response to these cross-pressures, agencies will more actively pursue statistical, administrative and legal approaches to responsible data dissemination. Recent developments in these approaches are discussed as they relate to improvements in database techniques, computer and analytical methodologies and legal and administrative arrangements for access to and protection of federal statistics.

*Key words and phrases:* Federal statistical system, data access, privacy, disclosure limitation, masking.

## 1. INTRODUCTION

As suggested by its subtitle, our purpose in this article is to consider what the not-too-distant future holds for an important issue facing the statistical community. Our thoughts are offered to advance a discussion of how we can better mediate the growing tension between confidentiality and data access. We recognize the simultaneous increase in the demand of the research community for access to federally collected microdata and the stake of respondents in the confidentiality of these records. With this understanding, the time has come for a new look at the issues raised, for example, in the 1978 Office of Federal Statistical Policy and Standards Report on Statistical Disclosure and Disclosure-Avoidance Techniques. Emblematic of this new look is the Panel on Confidentiality and Data Access of the Committee on National Statistics of the National Research Council and the Social Science Research Council, whose mandate is to provide the federal statistical community with recommendations for better managing this tension and whose work is ongoing.

The contemporary concern for the tension between access and confidentiality is also represented in the draft guidelines for statistical practices that

*George T. Duncan is Professor of Statistics at the School of Urban and Public Affairs, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213. Robert W. Pearson is Director of Corporate and Foundation Relations and Adjunct Assistant Professor of Political Science at Barnard College, 3009 Broadway, New York, New York 10027-6598.*

are currently under revision by the Office of Management and Budget. These guidelines seek to protect the confidence and trust of respondents in federal statistical agencies while gaining full value for the taxpayer from the information that these agencies collect.

## 2. CONTEMPORARY CONCERNS

Why should researchers have access to federally collected microdata? Providing researchers with access to government-collected microdata advances accepted public policy goals in a democratic society. U.S. taxpayers, for example, provided more than $3 billion in FY 1990 for statistical activities in 70 agencies of the Federal government, including $1.3 billion for the 1990 Decennial Census. By providing researchers with access to data, we permit (re)analysis that can shed light on questions—perhaps not envisioned when the data were originally collected—and that can verify results reported by the data collection agency or other analysts. Furthermore, such access stimulates new inquiries on important social, economic, and scientific questions; improves the quality of data by suggesting improved measurement and data collection methods; and provides information to improve forecasts and resource allocation (see, e.g., Flaherty, 1979; Fienberg, Martin and Straf, 1985). Such access helps provide facts that an electorate and its representatives need for informed decision making.

Widespread access to surveys such as the Panel Study of Income Dynamics and the National Longitudinal Surveys of Labor Market Experience, for example, have furthered our understanding of the dynamics of poverty, replacing longstanding beliefs about the permanence of poverty with knowledge about the extent to which poverty is both widespread and temporary for a large proportion of the American public (Duncan, 1984). Access to computerized criminal history files maintained by the FBI has permitted longitudinal studies of criminal careers, which have overturned some inferences drawn from previous cross-sectional studies of crime (Blumstein and Cohen, 1987).

In spite of the evident value of microdata dissemination, however, serious concerns about access to publicly collected microdata have been raised. Five factors give rise to this contemporary concern about the (re)identification of individual records.

**Identification is easier**. Sophisticated and more widely available computational and analytical technologies make it easier to breach the anonymity of subjects of publicly sponsored surveys and administrative records. Similarly, the increasing possibility of linking data files—both because more such files exist and because of automated record linkage procedures—make the possible disclosure of the identity of individual records in such linked files easier in principle.

**More microdata files exist**. Government and business have created and accumulated increasing numbers of microdata files.

**The consequences of disclosure are greater**. Those who collect these data are increasingly concerned that the technology and the detail of records will diminish the public's trust and cooperation with these data collection programs. As a result, the quality and, hence, usefulness of the data themselves will decline, as then must the ability of the agencies to fulfill their missions.

**The motivations for identification may have increased**. There are—in an increasingly information-based society—increasing incentives to gain advantage through intelligence-gathering activities, whether by government administrative agencies or by private organizations.

**Microdata files are harder to disguise**. Detailed microdata files—increasingly longitudinal in design—make the unique "signatures" of individual records increasingly difficult to disguise prior to their distribution without also degrading the scientific value of the data. It is now generally accepted—perhaps reluctantly by researchers requiring data—that the simple transformation of removing obvious identifiers or near identifiers (such as name, social security number, address or telephone number) is insufficient in many cases to hamper a serious data spy (see Paass, 1988), just as locking car doors does not deter a professional thief. Additional masking can deter all but the most determined spy, but some risk of disclosure must remain. Legislation governing access to data, however, is often written as if zero disclosure risk were required before data can be released. Taken literally, this would preclude researchers from access to microdata and deny society the benefits of the research.

The demand for microdata by social scientists is well established. For example, Arber's (1988) survey of British academics revealed a strong desire for the re-analysis of individual-level samples of census data that used academics' own hardware and software. A 1984 Census Bureau conference, for example, witnessed more than 100 economists expressing a need for a public use Longitudinal Establishment File (Govoni-Waite, 1985). Such demand is also revealed in the growth of such institutions as the Interuniversity Consortium of Political and Social Research and in testimony of social scientists in forums of the American Statistical Association, the Association of Public Data Users, and meetings of academics with federal statistical agencies. Indeed, one could easily propose a "law of data access," which includes the maxim that the greater the access to detailed records provided to social scientists, the greater will be their demand for more. The thirst of the scientific enterprise is necessarily and appropriately unquenchable.

But because of these five factors of contemporary concern, the supply side for microdata is hampered. Recent examples of the unmet need for microdata from one important federal statistical agency—the Bureau of the Census—include the following (Gates, 1988).

Researchers at Princeton University requested the exact date of birth on a microdata tape of the Survey of Income and Program Participation (SIPP) in order to study the Selective Service draft lotteries held in the United States in the 1970s. Because date of birth is available on many administrative record files and is an excellent match key, its inclusion on the tape would have increased the risk of identifying respondents to SIPP.

The Economic Research Service of the Department of Agriculture requested a file showing nonmetropolitan status of SIPP respondents in order to assess their economic well-being in terms of wealth, asset holdings and participation in government programs. The availability of information about such geographic units and the detail of individual data that were requested suggested the possibility of disclosure.

More generally, a number of social science research and public policy studies could be pursued if the present tension between access and confidentiality were better resolved.

Contextual data could be combined with data about individuals. For example, the National Longitudinal Surveys of Labor Market Experience (NLS) could link neighborhood and administrative data to the individual records of the NLS Youth cohort. Such linkages would enable the study of the processes by which persistent and concentrated urban poverty results in problems both for family processes and for the individual development of the nation's disadvantaged youth. Longitudinal, hierarchical data about students, classrooms and schools could be combined with data about the social structures of the communities and neighborhoods in which youth live and the street corners on which they play (and, for some, die), to better understand the way in which context mediates the relationship between individual abilities, academic achievement and employment outcomes.

The sponsorship of ongoing longitudinal surveys could be transferred from one agency to another as respondents age. The programmatic interests of several statistical agencies, for example, are tied to different stages in the life courses of people, but concerns about confidentiality have made it difficult for agencies to transfer responsibilities for data collection and analysis. For example, the Longitudinal Retirement History Survey has been of interest to the National Institute of Aging and the growing field of research and public policy concerning America's elderly population, but the transfer of responsibility for these data has been discouraged because of prohibitions on the release of these data and on their linkage to Social Security data, Medicare records and data from the National Death Index because of the possibility that such data could be identified by federal agencies who hold such information.

The latest scientific developments in analyzing very large spatial data bases and modeling complex spatial phenomena would be available. These developments, which could help achieve the goal of identifying and explaining human behavior at both the aggregate and individual levels require the use of refined geographic identifiers, which are not now generally available.

The concern by agencies for protecting the confidentiality of records is engendered by legal requirements, ethical issues involving actual and implicit commitments made to data respondents and practical worries about response rates to statistical surveys. An important part of any future data-disseminating program will be an adequate set of disclosure-limiting procedures that can be affected through various mixes of statistical, legal, administrative and ethical controls.

Our purpose in this article is to reflect on what the near-future holds for the mediation of concerns about data access and confidentiality. We draw on recent developments, and we paint a hopeful portrait of the future in part to help provide a target or goal—even if always moving—for better accommodating the increasing tension between data access and confidentiality.

In brief, our avowedly optimistic vision of the future looks like this:

- *Agencies will employ statistical masks that are effective yet faithful to the original data. Statistical methods for the analysis of masked data will be developed, cheaply available and easy to use.*
- *Electronic gatekeepers and monitors for the remote access to, and utilization of, computer databases will be widespread.*
- *Techniques for assessing the disclosure implications of record linkage and matching procedures will be further developed and routinely used in evaluating disclosure risks.*
- *Agencies will place more responsibility on researchers. Pledges, bonds and licensing contracts will become an increasingly explicit part of the*

*conditions under which researchers gain access to microdata.*

- *Legislation will recognize the need for research access and provide for sanctions for improper use of data, while recognizing the infeasibility of zero disclosure risks.*
- *Researchers' codes of conduct concerning disclosure will be further developed and widely discussed and will continue to be observed in practice.*
- *As an ethical prerogative, respondents will be better informed of the intended and potential research uses of the data they provide and will be apprised of the possibility, even if remote, of re-identification.*

Although this optimistic vision is feasible, it will require substantial effort. Otherwise, a bleaker vision of the future may look like this:

- *Agencies employ masks that make data difficult to analyze yet fail to deter data spies.*
- *Researchers are haphazardly denied access to federally collected databases, while data spies readily obtain personal information from private sources of information.*
- *Researchers, agencies and legislators spend considerable time wringing their hands about a seemingly chaotic and unprincipled and inequitable process of data access and data denial.*
- *Especially controversial data are held exclusively by federal agencies who fail to release information that may be embarrassing to government agencies, thus limiting our ability to understand these issues.*

In the remainder of this article, we will explore the optimistic version of these two simplistic visions and hope that, in pursuing its realization, we might avoid the pessimistic scenario. We begin with the statistical arena and look at statistical masks as a technique for disclosure limitation.

## 3. MASKING DATA AND DISCLOSURE LIMITATION TECHNOLOGIES

In safeguarding respondent privacy, data can be masked either by the respondent at the time of collection or by the agency at the time of release. In the former case, a technique for avoiding evasive answer bias is the "randomized response" technique introduced by Warner (1965) and discussed in the present context by Dalenius (1988). This technique is not widely used in survey practices, however, and has the disadvantage of limiting analysis of information collected by this technique to the reporting of univariate statistics. We focus instead on the second case of masking data at the time of release.

At present, many microdata files are released after the agency has masked the data to limit the possibility of disclosure. Typically, names and other identifying information are removed from them before being released for research use. Beyond such de-identification, the U.S. Bureau of the Census, for example, uses the release of sampled data as a disclosure-limiting device, a practice it began using when it provided public use microdata from the 1960 decennial census as a one-in-one-thousand sample file (Gates, 1988).

Data are typically held by an agency in a file represented by an $n$-by-$p$ matrix $X$. Each of the $n$ rows gives individual data on each of $p$ attributes. A file records many attributes of respondents, including some that are sensitive (e.g., income, assets or medical conditions). Some attributes are formal identifiers, such as name and social security number, which are removed in de-identification before release. Some attributes are "quasi-identifiers," in that their values are publicly known, such as sex and race.

The data may be masked through such methods as: (1) releasing only a sample of the data (subtracting rows from $X$); (2) including simulated data (adding rows to $X$); (3) blurring (fuzzing individual values in $X$ by random rounding, grouping, adding random error, etc.); (4) excluding certain attributes (removing columns of $X$); and (5) swapping (exchanging blocks of rows in a certain subset of columns of $X$).

The purpose of masking data is to make it more difficult for a data user to break the confidentiality of the database $X$. In the evocative language of Roberts (1986), such a user would be referred to as a statistical spy. However, because statistical purposes are correctly held as appropriate uses of the data, we instead refer to such a user as a *data spy*.

A careful consideration of the deterrence value of various masking methods is required if data custodians are to be convinced that microdata can be released under statistical controls. In addition, the potential of masked data for valid and informative statistical inference must be assessed and new methods of statistical analysis of masked data developed.

In examining the deterrence value of a particular transformation, the disclosure-limiting (DL) approach of Duncan and Lambert (1986) begins by modeling the decision problem of the data spy in inferring the value of a target $Y$ from the released $X$. This target value may be considered sensitive, as in a survey prompted by concerns about AIDS in which the value may indicate the number of sexual partners or the engagement in unorthodox (and in some states, illegal) sexual practices. Many other attributes may be considered sensitive in certain circumstances. Even the proverbial "known" attribute of one's age is sensitive for many people and can be a determinant of pension entitlements. Marital status is sensitive for some, as can be the number of children ever born to a respondent.

As discussed in Duncan and Lambert (1989), the specific piece of information—the target $Y$—sought by the data spy may refer to: (1) a specific respondent or (2) any respondent in the data base. Disclosure limitation methods such as using small sampling rates are effective against the first objective of a data spy but not the second. The second objective might be held by a data spy that sought to embarrass the statistical agency. For this latter purpose, finding *any* identifiable information about *any* respondent suffices. Small sampling rates are less effective against this strategy.

A measure of inferential disclosure risk (Duncan and Lambert, 1989) is the information in $X$ for inferring $Y$. The DL approach seeks to raise the price of inferring protected values from the released data so high that the spy will not take such actions. The intention is not simply to avoid having the spy make correct inferences. It is just as important that the spy refrain from making identifications altogether—whether correct or not—both because any purported inference can damage a data-disseminating agency and because luring the spy to incorrect inferences can typically only be achieved by releasing misleading data, which can undermine legitimate research. From a decision-theoretic point of view, DL approaches raise the Bayes risk of inference high enough so that the option of no inference is preferred. This philosophy yields the threshold rule for a statistical agency: Release the data if the Bayes risk to the data spy exceeds some threshold.

The nature of the inferences that a data spy may make from released microdata can vary substantially. Hence, disclosure can be conceptualized in various ways. Spruill (1983), Paass (1988) and Strudler, Oh and Scheuren (1986) equated disclosure with the identification of a respondent from a released file. Duncan and Lambert (1989) called this *identity disclosure.* Cox and Sande (1979) equated disclosure with obtaining reliable information about a respondent as a result of linking a record to the respondent—the *attribute disclosure* of Duncan and Lambert (1989). Dalenius (1977b) and the Subcommittee on Disclosure Avoidance Techniques (1978) equated disclosure with inferring new information about a respondent from the released data, even if no released record is associated with the respondent and the new information may itself be inexact—Duncan and Lambert's (1989) *inferential disclosure.* Palley and Simonoff (1986) equated disclosure with inferring certain characteristics of a population or a model; for example, the tax compliance model of the Internal Revenue Service. Duncan and Lambert (1989) refer to this type of disclosure as *population* or *model disclosure.*

Statutes regarding confidentiality are generally concerned with the probabilities of identity disclosure—a data spy using the information, perhaps in conjunction with collateral information, to identify a particular individual or institution within a record system. Statutes sanction this type of disclosure exclusively. Yet these other types of disclosure are possible under some conditions and could discredit the statistical agency if these distinctions are not maintained.

Regardless of the type of disclosure considered, a DL approach assesses the conditional distribution of a target value $Y$, given the masked data. The heuristic motivation behind this conditional distribution or predictive approach is evident: The intruder wants to use the information in the masked data $X$ to infer something about the sensitive target value $Y$. All probability distributions have the following interpretation: They are the subjective distributions of the intruder as they are perceived by the data disseminating agency.

To sharpen this discussion, we focus our attention on the use of *matrix masking* of the microdata file $X$. The data user is provided the masked microdata file $M = AXB + C$ and is not given the original data $X$. The matrix $A$, as a matrix of row operators, directly transforms the data records in $X$; so we call $A$ a record-transforming mask. The matrix $B$, as a matrix of column operators, directly transforms the data attributes in $X$; so we call $B$ an attribute-transforming mask. The matrix $C$ displaces $AXB$ by adding stochastic or systematic noise to the data; so we call $C$ a displacing mask. In general, the mask $(A, B, C)$ may depend on the particular values in $X$. That is, the mask components $A$, $B$ and $C$ are not necessarily just fixed matrices with constant elements or random matrices with elements that are independent of the values in $X$.

Generally, because the data must be analyzed, the data provider must also give the user either the complete specification of the mask $(A, B, C)$ or certain characteristics of it. It is an open question of disclosure-limitation methodology as to how much information should be given the data user about the mask in a particular context (Wolf, 1988). Clearly in the case of $M = AX$, for example, $A$ cannot be specified if $A^{-1}$ exists, that is, if the privacy transformation is reversible, as discussed in Dalenius (1977a).

Matrix masks are powerful—and thus likely to be increasingly used by statistical agencies—because they encompass many commonly proposed disclosure-limitation methods. We illustrate this first with record transforming masks $A$, second with attribute transforming masks $B$ and third with displacing masks $C$. Some of these procedures are discussed in McGuckin and Nguyen (1988a) and in Dalenius (1988).

### Record Transforming Masks

By changing the form of the record transforming mask $A$—even with $B$ an identity matrix and $C$ a zero matrix—we can represent some currently proposed disclosure-limitation techniques, such as the following.

**Aggregation across records**. For example, averaging all attributes over three similar records. Here $A$ depends on $X$, because of the use of "similar" records.

**Suppression of certain records**. For example, suppression of records having extreme values on some attributes or suppression of records from small identifiable geographic units. Here again the transforming mask is a function of the data file $X$.

**Release statistics for regression**. Take $A = X'$, then $M = X'X$ is sufficient for ordinary least-squares regression of any attribute in $X$ on any subset of other attributes. (This point was suggested to us by Steven Klepper.)

We can also consider a random record transforming mask in which the matrix $A$ has stochastic elements. Special cases of interest include the following.

**Sampling**. In sampling $r$ rows of $X$, the matrix $A$ has 0-1 random entries with a single 1 in each of $r$ rows.

**Multiplication of records by random noise**. With the matrix $A$ diagonal, each record is multiplied by a random variable.

### Attribute Transforming Masks

By changing the form of the attribute transforming mask $B$, we can represent the following DL procedures.

**Aggregation across certain attributes**. For example, the release of total income, rather than the (disaggregated) release of salary income, business income, interest income, etc.

**Suppression of certain attributes**. For example, some attributes—such as identifiers or medical conditions such as mental health or HIV infection indicators—may be suppressed.

**Multiplication of attributes by random noise**. With the matrix $B$ diagonal, each attribute is multiplied by a random variable.

### Displacing Masks

In the case of displacing masks (the matrices $A$ and $B$ are identities), adding $C$ yields the following DL techniques:

**Addition of random noise**. Adding a random variable to each entry.

**Addition of deterministic noise**. Adding a specified quantity to each entry.

Often, implemented procedures involve a combination of DL procedures. See, for example, Kim (1986) for a Census Bureau application to the Continuous Longitudinal Manpower Survey, which was conducted for the Bureau of Labor Statistics to evaluate the effectiveness of the Comprehensive Employment and Training Act (CETA) of 1973. The public use files contain earnings data matched to Social Security Administration administrative records. The masking technique involved both the addition of random noise and data transformation. In these cases, the transforming masks $A$ and $B$ are not identity matrices and the displacing mask $C$ is not the zero matrix.

Given the richness of matrix masks, it is reasonable to ask what commonly used (or proposed) DL procedures are not matrix masks? Such examples would include: (1) *attribute-specific aggregation over records* (release of some attribute values unmasked, but aggregating other attribute values—say releasing only averages of interest income for similar records); (2) *data swapping* (release of records with some, but not all, attribute fields interchanged); (3) *multiplication by random noise* (multiplying each element of $X$ by mutually independent random variables is not a matrix multiplication or addition); (4) *random rounding* (rounding each entry to a certain base); (5) *grouping* (condensing categories for some attributes); and (6) *truncating* (truncating distributions of certain attributes).

Generally, ad hoc arguments have been used to devise disclosure-limitation procedures and to evaluate them in terms of disclosure risk and data utility. Studies to date suggest that particular implementations can result in significant differences between the information provided by the masked data and that available from the original file (see, for example, Wolf, 1988, for an assessment of surrogate microaggregate records). This suggests that a more general analysis based on a systematic approach to masking is desirable.

In disclosure limitation, we seek a mask that leaves the maximum information about $X$, while preserving its confidentiality. In a specific application, which requires a (vector) statistic $T(X)$, the mask $M$ should minimize the difference between $T(M(X))$ and $T(X)$, while maximizing the difficulty of a data spy to infer the target $Y$ from $M(X)$. As a generally useful approach, this suggests choosing a mask $(M)$ to minimize the conditional variance of $X$ given $M$ while maximizing the conditional variance of $Y$ given $M$. This notion of constrained optimization can be considered consistent with what is reported to be Census Bureau policy: "In practice the Census Bureau has taken disclosure protection as a binding constraint and provided as much data to the public as is possible within this constraint" (McGuckin and Nguyen, 1988b). This approach has no value when all attributes of all persons are sensitive; that is, when the entire $X$ matrix is sensitive and the target $Y$ equals $X$.

While these and other disclosure-limitation techniques promise to help mediate the tension between access and confidentiality, their implementation carries with it a somewhat paradoxical danger that the more sophisticated the masking technique, the less accessible the data and their analysis will be to many social scientists and policy makers. This danger arises because the researcher must analyze the data in the masked form $M$ rather than in the original form $X$. In the case of masking through sampling, standard tools are appropriate. But the addition of noise, for example, presents measurement error or errors-in-variables problems for the user analyzing the masked data (see, e.g., Sullivan and Fuller, 1989). Social scientists will require new training in the use of such masked data, and special care will be required in interpreting masked data. Further, in some cases, new statistical procedures will be required for analyzing masked data.

Some researchers have begun to address these issues. Kamlet, Klepper and Frank (1985), for example, analyze the 1980 National Health Interview Survey (NHIS) in which several averages from aggregation are reported rather than individual-level data because of confidentiality restrictions. (NHIS is a stratified cluster sample of approximately 25,000 U.S. households that is conducted by the National Center for Health Statistics.) Typically, analysts of such data simply use the associated group-level information instead of the (unobserved) individual-level data. As Kamlet, Klepper and Frank note, however, this practice can introduce measurement error in an explanatory variable, which can produce inconsistent estimates and regression coefficients of the wrong sign. Kamlet and Klepper (1985) demonstrate how consistent estimators can be computed in certain special cases.

## 4. RESTRICTED OR CONTROLLED ACCESS TO DATA

Masking data prior to their release as public use microdata files will not in all cases be sufficient (or necessary) to protect data from the data spy who lurks beyond the walls of the data-providing agency. Additional or alternative lines of protection will be required in many instances. These protections themselves represent a wide range of methods, from electronic gatekeepers and monitors to contractual licensing agreements that provide penalties for the misuse of data.

### Electronic Gatekeepers and Monitors

In some cases, access to data by a researcher will be controlled by an intermediary—or "gatekeeper" —as contrasted to or in addition to masking microdata files prior to their release. Increasingly, researchers will want to access large-scale statistical data sets through computerized telecommunications networks. Because of storage and maintenance efficiency, certain comprehensive data sets will more frequently be consolidated. Networks allow the researcher remote access, avoiding trips to the site, and can permit the use of the researcher's own software in the analysis of the data.

An important current example of such an arrangement is the Luxembourg Income Study (LIS) (Rainwater and Smeeding, 1988). Sophisticated microdata sets that contain comprehensive measures of income and economic well-being for many developed countries are centrally stored at the Center for Population, Poverty, and Policy Studies (CEPS) in Luxembourg. Because of the dual considerations of the cost of international researchers directly accessing the microdata in Luxembourg and confidentiality concerns about the release of public use data files, computer network access has been implemented through a major international network (BITNET), the European Academic Research Network (EARN) and a network spanning Canada (NETNORTH). In the form that was implemented in September 1989, LIS isolates the researcher from the data file. Requests are submitted to an electronic gatekeeper that checks that: (1) the user has been authorized to use the database and (2) the requested SPSSX run does not contain commands that could result in "stealing" individual cases of microdata. Jobs submitted from remote sites that fail security or syntax checks are sent to a special machine for review by the technical staff (Luxembourg Income Study Newsletter, July 1989).

In the Luxembourg Income Study, the obligations of researchers to confidentiality are emphasized through the fact that all LIS output contains the following message:

> Use of the data in the Luxembourg Income Study database is governed by regulations which do not allow copying or further distribution of the survey microdata. Anyone violating these regultions will lose all privileges to the database and may be subject to prosecution

> under the law. In addition, any attempt to circumvent the LIS processing system or unauthorized entry into the computers of the Centre Informatique de L'Etat of Luxembourg will result in prosecution (Luxembourg Income Study Newsletter, July 1989).

As organizations increasingly employ distributed database systems, new concerns about data integrity and security in information networks have arisen. Authorization policies and implementation strategies of trusted networks must accommodate the varying levels of security at the network nodes—including the class of home computers with dial-up potential—so that sensitive information can be processed.

The initial focus of network security has been the problem of controlling access to systems and files at a macro level. While necessary, such access control—say, by passwords—is not sufficient to protect the privacy and integrity of sensitive information. Network security must also encompass utilization control, which can be thought of as access control at a micro level.

By analogy, the guard at the art museum's gate qualifies entrants (thereby controlling access to the museum), but additional security measures are needed in utilization of the museum to prevent theft and vandalism (thereby controlling access to the individual works of art). Developments in computer science promise to provide mechanisms for these room monitors as well as gatekeepers.

Increasingly, organizations are establishing statistical databases that reside on computers and contain confidential data or, implicitly, relationships that are of a sensitive nature. Blue Cross and Blue Shield of Massachusetts, for example, has established the Provider Terminal Network, which allows physicians and hospitals to directly verify a patient's status and eligibility. More generally, the increased amount of confidential data transmitted over networks has prompted the TeleCommunications Association and large network users to appeal to the FCC to determine what network data are considered proprietary by customers. Further, the Computer Security Act of 1987 requires that civilian agencies identify systems containing sensitive information and develop a security plan for each sensitive system. With their proliferation, the data held in these networked systems will become of increasing interest to researchers.

Macro-level access control techniques prevent unauthorized access to networks by verifying a user's identity prior to allowing the user access to the host or the network. There are many techniques for making access to a network secure, such as authentication, passwords and encryption. Most access control techniques are not fully relevant when a user has legitimate access to certain information, say, certain statistical aggregates, but does not have legitimate access to certain other information, say, medical, sales or salary information that is identifiable to a particular individual. Limiting queries to statistical aggregates is insufficient because a series of such queries can readily identify individual information (see, e.g., Ahituv, Lapid and Neumann, 1988). Current, as well as post facto, monitoring of repeated access to aggregates may require the comparisons of the ranks of matrices that, in practice, are intractably large (Fellegi, 1972).

More sophisticated authorization rules will be needed to determine what users can do or see. While some formal theory has been developed for this purpose (see, e.g., Landwehr, 1981; Denning, 1982), current techniques for utilization control are fairly rudimentary. For example, audit trails operate only ex post facto in establishing what a user has done. Multilevel passwords for applications and records provide only limited flexibility in controlling utilization.

Secure databases permit users to query the database according to certain authorization rules. A database has been compromised when a database spy has identified a confidential data record or identified a restricted relationship. Alternative disclosure limitation techniques should be pursued in this context: (1) limiting the query set, (2) limiting the intersection of query sets, (3) random sample queries, (4) partitioning the database and (5) perturbing data values (Shosani, 1982). These techniques warrant systematic investigation so that networked database systems can achieve their full potential for the researcher.

### Legal, Administrative and Contractual Arrangements for Limiting and Controlling Access and Penalizing Misuse

Research access to data is controlled through a variety of regulations and laws. Improvements in computer technology motivate many of these changes. Often, the development of legal controls lags behind changes in technology, however. Courts, for example, have been slow to recognize the substantive difference between manual records and computer records. And some have argued that the Privacy Act of 1974 has been rendered obsolete by technological developments in the years since the law was passed (see Dean, 1986).

Some regulatory attempts to restrict access would, as in a 1986 National Security Council directive, limit the use of commercial data bases.

These attempts were aborted in 1987 under pressure from the American Civil Liberties Union and the Information Industry Association. "Before these computerized information banks were created, such technical reports were scattered in hundreds of arcane journals and libraries. Now the data-base companies collect millions of documents and let customers comb through them in minutes by computer" (Davis, 1987). In Great Britain, the Data Protection Act of 1984 regulates the storage and processing by computers of data about living individuals. As the Act applies to data held for statistical or research purposes, the Royal Statistical Society formed an ad hoc study group to monitor its impact.

Legislation governing access to data varies from one agency to the next in the United States, and in some cases varies within an agency (e.g., Titles 13 and 15 prescribe different treatments for data collected by the Bureau of the Census). The future is likely to retain this diversity, but some convergence in laws and practices may occur as issues of confidentiality and access arise with each reauthorization of agencies as they begin to draw on the experiences of others in designing guidelines for access and confidentiality.

For example, the National Center for Education Statistics was recently required by its authorizing legislation to design such guidelines. These draft guidelines drew on existing legal and administrative models at the Bureau of the Census and the Department of Justice in providing substantial sanctions against misuse and assurances against individual records being subpoenaed and in providing for tests of the ability to identify records through the use of readily available collateral information about states, districts and schools. Draft NCES guidelines also provide for a microdata review panel to consider whether and how specific data sets are to be made available for research. The guidelines further suggest that regional centers be established where analysis of data may be conducted by researchers as specially sworn employees of the Center under supervised and monitored conditions. Some rationalization of such practices across agencies may also result from the activities of the panel of the Committee on National Statistics of the National Research Council and the Social Science Research Council, which we noted at the beginning of this article.

Administrative arrangements for controlling access to microdata that have not been released for public use include the extension of legal responsibilities and sanctions to the outside analyst. The Privacy Act provides for the use of specially sworn employees to analyze such data, and the Bureau of the Census has used this provision in the law through certain fellowship programs in which it participates to provide access to data that could not be otherwise provided because of the possibilities of reidentifying individual records.

Such arrangements have had a mixed, and not as yet fully assessed, record. They appear to discourage use by researchers who are increasingly taking advantage of the reduced costs of personal computers to conduct flexible, sophisticated, custom and inexpensive analyses of all but the very largest data sets typically available in the social sciences.

Specially sworn employees have also been required to relocate to the principal office of the agency (e.g., Suitland, Maryland), where they must use computer hardware and software configurations that may differ from their own. The privileges and responsibilities that accrue to sworn employees are usually temporary. They confer rights and responsibilities for the duration of the grant or fellowship and require (in the case of the Bureau of the Census) that the research be directly applicable to the mission of the agency.

While our understanding of the controlled access experiences is primarily anecdotal, we would conclude that such arrangements have: frustrated researchers; added costs in the amount of time required to learn the strengths and weaknesses of the data and the often different computing environments in which data can be analyzed; added considerably to the time required to complete the research (e.g., Levin and Stephan, 1988, estimate that arrangements that permitted their analysis of a linked data file of citations and characteristics of male scientists added 18 months to the time required to complete the study); discouraged uses by researchers capable of analyzing the data and in creating and sharing information about its quality; and required resources on the part of the agency (e.g., office space, personnel required to review output and monitor use of files, etc.) that were often difficult to secure or spare.

Currently, pilot programs are being conducted in which such access is provided to specially sworn employees who work under supervision in more widely distributed regional centers or offices. While experience with such regional arrangements will permit emendations and refinements of these practices, they will remain able to serve only a small fraction of the research community. Further, they will increasingly come into conflict with the competing needs for open access to data which are suggested by recent cases of scientific fraud and the desire for reanalysis.

Licensing agreements, on the other hand, promise to become increasingly used as a means of making explicit the contractual and ethical obligations of the researcher to care for the protection of the records, while providing sanctions against misuse. For example, Ohio State University provides a special "geocode data tape" containing county data, college identifiers and some administrative data for the National Longitudinal Survey of Labor Market Experience Youth Cohort (NLSY) under license to other institutions. The Panel Study of Income Dynamics (PSID) is appending census tract information to its records and will release a public use file with this information. Access to this file requires universities to sign a detailed license agreement and provide a deposit of $1,000, which will be returned when the data are returned and a full accounting of their use is provided.

More explicitly and widely applied sanctions against the misuse of statistical records will unlikely change the behavior of most researchers. They have few incentives and limited resources to violate the anonymity of individual records. Indeed, this lack of incentives to identify individual records remains the most powerful deterrent to this type of misuse. In addition, such sanctions may be difficult to enforce, although not impossible.

Quite apart from its enforceability, however, penalties signal shared norms of behavior whose violation or transgression is taboo. Such functions are not trivial. Social organizations assign priorities to such norms and values by publicly providing rewards for compliance and penalties for violation. Those norms that have carrots and sticks are more important than those that do not.

The threat of sanctions (in combination with professional codes of conduct) influence the behavior of federal statistical officers in encouraging them to care about the way in which record systems are managed. To have in place a system of penalties that can be invoked in the rare case of misuse is a useful tool. It helps contribute to an orderly civil society and may help restore or protect public faith in the system of norms and contracts by which a well-functioning society works. Even though many criminals escape detection, incarceration, trial and imprisonment, we are still pleased to know that sanctions exist for criminal behavior and can hope that they would serve us justly if we were to require them or be an object of their administration.

In part, the effort to add the threat of sanctions to the calculus of end users involves a political strategy of sharing obligations and responsibilities, which now asymmetrically apply only to federal statistical officers or their sworn agents. This sharing seems on the face of it a reasonable request to make of the social scientific community who continually demand greater access to data. Agreeing to abide by similar standards of stewardship that are required of those who provide the data removes a crutch that can be used to retard access.

Previous experience with special licensing arrangements clearly suggests that some care should be devoted to their implementation and assessment, however. Statistical Policy Working Paper 2 (Report on Statistical Disclosure and Disclosure-Avoidance Techniques) issued by the Federal Committee on Statistical Methodology of the then Office of Federal Statistical Policy and Standards, prepared the last comprehensive examination of similar arrangements among federal statistical agencies in 1978. Concerning the use of the 1960 Census public use samples, which then required purchasers to sign an agreement that prohibited any dissemination of the samples to a third party without written authorization from the Bureau, the report concluded (page 31):

> By 1969 the Bureau had sold over sixty-five copies of the files, but had received only a handful of publications and requests to approve copying the files for a third party. At the same time many other publications based on the public-use sample data were found, few of which contained the required disclaimer, and it was estimated that the files were available in over 200 institutions. ... [T]he necessity of more complete arrangements with purchasers of restricted use files, include periodic follow-up, and denying access to researchers who are not able to control completely the handling of the data files in question within their institutions.

In its conclusion, the report, however, also noted:

> While we have found some examples of what we consider to be unacceptable statistical disclosures, we have not been able, in spite of a fairly systematic effort, to locate a single instance in which an individual (natural person) alleged that he or she was harmed or might be harmed in any way by statistical disclosure resulting from data released by Federal agencies.

Given how quickly data files were informally disseminated in the 1960s, we would currently expect—with the lower costs of computing and the widespread availability of networks—that files can be made more rapidly available, without adequate controls, in thousands of institutions. Furthermore, the fact that no one has been harmed by the release of public use files to date is, in itself, not fully reassuring. The potential for harm to a respondent and a statistical agency remains, and it appears to strongly influence the decisions of federal statistical agencies who fear even one case arising that would require them to stand before a court of inquiry and argue persuasively they were careful stewards of data.

### Explicitly Requesting Permission to Provide Research Access to Records While Recognizing the Possibility of Disclosure

Recently, federal agencies have turned their attention toward improving the manner in which they inform respondents of the prospective uses of the data. While future uses cannot always be anticipated, the types of uses that are now easily conducted and increasingly called for suggest that federal agencies could be more explicit about anticipated uses beyond the simple statement that they will be used only for statistical purposes, as most federally sponsored surveys now report.

In one pilot study, the Bureau of the Census examined the consequences of more elaborate forms of informed consent that explicitly requested respondents' permission to permit access to data that are to be linked to administrative records. This pilot study was based on approximately 400 respondents to the 1989 Survey of Work Experience of Mature Women, who had been interviewed regularly since the mid-1960s. At the conclusion of the 1989 pilot interview, a subset of respondents were asked to read a statement that requested their permission to add to their research records their earnings and retirement and disability benefits records from the Social Security Administration and Internal Revenue Service, and Medicare and other medical care benefits from the Health Care Financing Administration. The statement that respondents read, and on which their signature was required, explicitly described the research uses to be made of such data. It also noted that, in adding such data, the possibility of a government agency re-identifying their records would increase. By signing such statements, respondents were authorizing, in a way far more elaborate than typical surveys, a specific use of the data they provided and in an important sense were exercising control over their data beyond turning them over to a paternalistic agency for safekeeping.

Of the 336 respondents who were asked to consent to such uses and linkages, 66 percent agreed. The Bureau debriefed interviewers and analyzed the differences between those who agreed and those who refused. There appeared no clear or dramatic differences in the degree of consent/refusal by the age, race or levels of education of the respondents, although the analysis revealed differences ranging from 3 to 8 percentage points (the younger, black and less highly educated tended to provide somewhat higher levels of consent). The responses, however, differed markedly by the size of place and by specific cities. Only 49% of those respondents from large cities (32 of 65 respondents) consented to the use and linkage of such data. Relatively low consent levels by respondents in specific cities such as New York (38%; 8 of 21 respondents) and Chicago (47%; 8 of 17 respondents) suggest that people living in such places are considerably more reluctant or fearful of the requested uses. Moreover, a 34% overall refusal rate raises concerns about a selection bias in the analysis of such linked data. The Bureau analysts themselves noted that the consent statement was confusing, and they proposed, in a similar request being made of the 1990 resurvey of a comparable cohort of mature men, to explain more fully the purposes of the request by using simpler language (Liebrecht and Smilay, 1990).

Other evidence suggests that under certain circumstances more respondents than those noted above are willing to permit research uses of the information they provide, even if the possibility of re-identification increases. This evidence comes from an inadvertent quasi-experiment involving a survey of Ph.D.s. Fully identifiable files of doctorates awarded on or after July 1, 1980, were first made available to all the government sponsors of the Survey of Earned Doctorates (SED) in 1981. To reflect this change, the confidentiality statement for the SED was modified in order to inform respondents of this change. However, several universities unwittingly distributed older questionnaire forms to their graduates, which did not state that identifiable files would be provided to sponsoring agencies. As researchers themselves, these respondents are presumably sympathetic with the research activities of the sponsoring agencies and understand the trade-off between the public good of social research and personal privacy.

To rectify this mistake, the National Research Council on four separate occasions wrote "old form" respondents a letter that both informed them of the mistake and permitted them to prohibit their records from being released to the sponsoring agencies (Coyle, 1988). In all, less than 0.5% of the approximately 6,000 people to whom the passive waiver consent letter was addressed requested that their records not be transferred to the sponsoring agencies.

Surely, in other circumstances, the authorization rate may be expected to be lower as respondents find it difficult (or are weakly motivated) to comprehend consent statements. Many household surveys compound this problem by collecting information from a single respondent about all members of the household. Our point in this very brief review of two recent experiences with consent is to suggest the promise of such a device for securing the cooperation of respondents to a more extended use of data than federal agencies have hertofore attempted or have been able to make.

In summary, recent developments suggest that increasing attention will be devoted to the theory and practice of informed consent as it relates to providing access to surveys and linked data files. The lessons drawn from the considerable attention to these issues in biomedicine during the last 10 years may be increasingly imported and applied to federal surveys. Agencies should be encouraged to draft informed consent agreements for respondents that assure that their privacy is protected, that response rates are not lessened and that legitimate research use of the data is authorized by respondents who are asked to consent to plans to use the data for research purposes. Assurances of these outcomes are likely to be made with greater confidence, however, only if federal agencies embark on a program of pilot studies that will empirically assess these outcomes.

Many lengthy and ongoing longitudinal studies did not fully anticipate all the uses to which such data could be put and, therefore, have occasionally failed to properly or fully inform respondents of such uses. Insofar as these studies continue, or addresses and locating information about sample respondents is current, agencies can (we may even argue, are obligated to) return to respondents (or their guardians) to renegotiate informed consent agreements concerning these more elaborate uses. We would predict that this practice would become more frequent, although in some cases difficult issues concerning the consent of guardians will arise, and continued monitoring of the best prevailing practices for securing informed consent will be required.

Another possibility in this area is to develop compensation procedures to pay reparations to individuals whose data have been disclosed. Such procedures may help increase the willingness of individuals to grant permission to match and link records from different sources, although we know of no evidence that addresses this mechanism directly.

## 5. CONCLUSION: AGENCIES AND RESEARCHERS AS DATA STEWARDS

Empowered by exponentially improving computer technology, researchers will have access to larger and more detailed databases in the future. This emerging capability provides an exciting opportunity to better employ factual evidence in developing public policy. As improved computer technologies increase both the value of data and its potential for compromise, however, statistical agencies and researchers will (must) increasingly assume the role of data stewards. As in the biblical parable, the best steward is one who ensures effective use of the data, not the one who protects it against any risk by hiding it, unused.

In exchange for such access, and as a symbol of their willingness to assume the role of steward, researchers will increasingly gain access to detailed records through masked data records and/or through licensing agreements that clearly state their responsibilities and liabilities. Attempts to establish personal identities of respondents will be explicitly proscribed as a condition for access to data and as part of the code of conduct professional research associations will take on in years ahead. Researchers will be required to protect the confidentiality of their data against outside threat and will be provided with legislative protection from subpoena of these records for the purpose of identifying individual subjects.

Statistical agencies, who currently bear the sole legal responsibility for protecting the confidentiality of records, will add to their concerns an affirmative obligation to conduct active review of the uses made of research records when there is some risk of disclosure. To fail to take these steps will diminish the prospects for achieving our benign forecast of the future.

We proposed above as a possible maxim in a "law of data access" that the greater the access to detailed records provided to social scientists, the greater would be their demand for more. This "law" could also contain another maxim, however, that suggests that the use of data is independent of the importance of the questions that the data can answer. In other words, the formulation of questions and the search for their answers will be sought often only where the light shines (to paraphrase a time-worn story of the drunk looking for his car keys under a street lamp because the light was shining there). We cannot be assured that we will find our car keys by venturing out into the unlit expanses of our current ignorance through providing continued and greater access to federal statistics, but the stakes are too high to be either foolishly adventurous or to limit our search too narrowly.

## ACKNOWLEDGMENTS

The authors gratefully acknowledge the suggestions of the Executive Editor and a referee, plus the helpful comments of Martin David, Daniel

Feenberg, Mark Kamlet, Steven Klepper and Diane Lambert. The work of George T. Duncan was partially supported under NSF Grant SES 89-10153. The work of Robert W. Pearson was partially supported under NSF Grant SES 89-12659 and grants from the Bureau of Labor Statistics, Bureau of the Census and the Statistics of Income Division of the Internal Revenue Service to the Social Science Research Council. None of these agencies necessarily shares the perspectives reflected in this article, nor do they bear any responsibility for its conclusions. An earlier version of this article was presented at the Annual Meeting of the American Statistical Association, August 1989.

## REFERENCES

Ahituv, N., Lapid, Y. and Neumann, S. (1988). Protecting statistical databases against retrieval of private information. *Computers and Security* **7** 59-63.

Arber, S. (1988). Anonymized samples of 1991 Census data. Presented at Access to the Census: Anonymized Samples and their Alternatives. Seminar sponsored by the Royal Statistical Society, Social Statistics Section, and the Social Research Association. London, England, November 15.

Blumstein, A. and Cohen, J. (1987). Characterizing criminal careers. *Science* 28 August **237** 985-991.

Boruch, R. and Cecil, J. (1979). Report from the United States: Emerging data protection and the social sciences' need for access to data. In *Data Protection and Social Science Research* (E. Mochmann and P. Muller, eds.) 104-128. Springer, New York.

Campbell, D. T., Boruch, R. F., Schwartz, R. D. and Steinberg, J. (1977). Confidentiality-preserving modes of access to files and interfile exchange for useful statistical analysis. *Evaluation Quarterly* **1** 269-299.

Cox, L. H. (1980). Suppression methodology and statistical disclosure control. *J. Amer. Statist. Assoc.* **75** 377-385.

Cox, L. H. and Sande, G. (1979). Techniques for preserving statistical confidentiality. In *Proceedings of the 42nd Meeting of the International Statistical Institute* **3** 499-512. Manila Philippine Organizing Committee, International Statistical Institute.

Coyle, S. (1988). Memorandum to Tom Jabine, November 17, in response to a Workshop on Confidentiality of and Access to Doctorate Records, November 3-5, 1988, sponsored by the Committee on National Statistics of the National Research Council and the Social Science Research Council.

Curran, W. J. (1986). Protecting confidentiality in epidemiological investigations by the Centers for Disease Control. *New England Journal of Medicine* **314** 1027-1028.

Dalenius, T. (1977a). Privacy transformations for statistical information systems. *J. Statist. Plann. Inference* **1** 73-86.

Dalenius, T. (1977b). Towards a methodology for statistical disclosure control. *Statistik Tidskrift* **5** 429-444.

Dalenius, T. (1985). Privacy and confidentiality in censuses and surveys. Presented at the Annual Meeting of the American Statistical Association.

Dalenius, T. (1988). Controlling invasion of privacy in surveys. Monograph. Department of Development and Research. Statistical Research Unit, Statistics Sweden. Stockholm, Sweden.

Dalenius, T. and Reiss, S. P. (1982). Data swapping: A technique for disclosure control. *J. Statist. Plann. Inference* **6** 73-85.

Davis, B. (1987). Federal agencies press database firms to curb access to 'sensitive' information. *The Wall Street Journal* February 5.

Dean, C. (1986). Computer use for news raises legal questions. *New York Times* September 29, A-12.

DeMillo, R. A., Dobkin, D. P. and Lipton, R. J. (1977). Even databases that lie can be compromised. *IEEE Trans. Software Eng.* **4** 73-75.

Denning, D. E. (1982). *Cryptography and Data Security*. Addison-Wesley, Reading, Mass.

Duncan, G. J. (1984). *Years of Poverty, Years of Plenty*. Institute for Social Research, Univ. Michigan, Ann Arbor.

Duncan, G. T. and Lambert, D. (1986). Disclosure-limited data dissemination (with discussion). *J. Amer. Statist. Assoc.* **81** 10-28.

Duncan, G. T. and Lambert, D. (1989). The risk of disclosure for microdata. *Journal of Business and Economic Statistics* **7** 207-217.

Fellegi, I. P. (1972). On the question of statistical confidentiality. *J. Amer. Statist. Assoc.* **67** 7-18.

Fellegi, I. P. and Phillips, J. L. (1974). Statistical confidentiality: some theory and applications to data dissemination. *Annals of Economic and Social Measurement* **3** 399-409.

Fernandez, E. B., Summers, R. C. and Wood, C. (1981). *Database Security and Integrity*. Addison-Wesley, Reading, Mass.

Fienberg, S. E., Martin, M. E. and Straf, M. L. (eds.) (1985). *Sharing Research Data*. National Academy Press, Washington, D.C.

Flaherty, D. H. (1979). *Privacy and Government Data Banks: An International Perspective*. Mansell, London.

Gates, G. W. (1988). Census Bureau microdata: providing useful research data while protecting the anonymity of respondents. In *Proceeding of the Section on Social Statistics* 235-240. Amer. Statist. Assoc., Alexandria, Va.

Govoni, J. P. and Waite, P. J. (1985). Development of a public use file for manufacturing. In *Proceedings of the Section on Business and Economic Statistics Section* 300-302. Amer. Statist. Assoc., Alexandria, Va.

Greenberg, B. (1988). Disclosure avoidance research for economic data. Presented to the Joint Advisory Committee Meeting. October 13-14, Oxon Hill, Md.

Kamlet, M. S. and Klepper, S. (1985). Mixing individual and group-level data. Working Paper. Dept. Social Sciences, Carnegie Mellon Univ.

Kamlet, M. S., Klepper, S. and Frank, R. G. (1985). Mixing micro and macro data: statistical issues and implication for data collection and reporting. In *Proceedings of the 1983 Public Health Conference on Records and Statistics*. Department of Health and Human Services, Hyattsville, Md.

Keller-McNulty, S., Unger, E. A. and McNulty, M. S. (1989). The protection of confidential data. Paper presented at the 21st Symposium on the Interface: Computing Science and Statistics, April 9-12, Orlando, Fla.

Kim, J. (1986). A method for limiting disclosure in microdata based on random noise and transformation. In *Proceedings of the Section on Survey Research Methods* 370-374. Amer. Statist. Assoc., Alexandria, Va.

Landwehr, C. E. (1981). Formal models for computer security. *ACM Computing Surveys*. **13** 247-278. [Reprinted in 1984 in *Advances in Computer System Security* (R. Turn, ed.) **2**. Artech House, Dedham, Mass.]

Leiss, E. L. (1982). *Principles of Data Security*. Plenum Press, New York.

Liebrecht, L. F. and Smilay, S. B. (1990). A memorandum for

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 dissemination" 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.*