Journal of Official Statistics
Vol. 10. No. 2, 1994, pp. 115–132
© Statistics Sweden

Conflicts Between the Needs for Access to Statistical Information and Demands for Confidentiality

Stephen E. Fienberg[1]

Abstract: With the growth of computer-based government records and the continued collection of statistical data for research, especially in the social sciences, there has been a concomitant growth in the desire to access statistical information by government, industry, and university-based researchers. Moreover, as a result of modern computer technology and ever-expanding computer networks, the costs of data acquisition and transfer continue to drop, and the desirability of access to statistical information collected by others increases. While government statistical agencies and survey researchers have always been concerned about the need to preserve the confidentiality of respondents to ensure the quality of statistical data, these concerns have been heightened by the decline in response rates for censuses and surveys over the past two decades. This paper examines the seeming conflicts between the two perspectives of data access and confidentiality protection and briefly outlines some of the issues involved from the perspectives of governments, statistical agencies, other large-scale gatherers of data, and individual researchers.

Key words: Cell suppression; confidentiality; intruders; masking; microdata access; probabilistic data disclosure.

1. Introduction

The past three decades have witnessed not only the dramatic growth of computer-based government records but also the increased focus on the collection of statistical data for research, especially in the social sciences. Not surprisingly, there has been a concomitant growth in the desire to access statistical information by government, industry, and university-based researchers, in order to take advantage of data already collected and stored in computer-readable form. Indeed, a new information industry has arisen around the use of government information merged with private statistical records and then repackaged and sold to both business and government. And, at least in the United States, the federal government has advocated the private dissemination of government statistical data as a mechanism to

[1] Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A.
Acknowledgements: An earlier version of this paper was presented at the *International Seminar on Statistical Confidentiality*, organized by EUROSTAT and ISI, held on September 8–10, 1992, Dublin, Ireland. Its preparation was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada to York University, Toronto, Canada. I am indebted to George Duncan, William Kruskal, an Associate Editor and the referees for comments on an earlier draft of this paper and to Luigi Biggeri, Wouter Keller, Nancy Kirkendall, Udi Makov, and Chris Skinner for helpful conversations.

help defray the costs of data collection. The demand for statistical data comes not only from the traditional research and policy constituencies, but also from banks, law firms, marketing organizations, and even private investigators and journalists. The sellers of U.S. government data thrive by providing access to individually identifiable files in various administrative data bases such as those containing drivers' license records, real estate transactions, and lists of individuals who are delinquent in paying their taxes (e.g., see Rothfeder 1992).

Moreover, during the past decade we entered a new era of computing and telecommunications whose effect is only now being realized. As a result of modern distributed computing environments, mass storage devices, and ever-expanding worldwide computer networks, the cost of data acquisition and data transfer continues to drop. Of course, with the increase in computational power available to most statistical analysts and the ability to analyze larger and larger data sets with innovative methodologies comes the prospect of bigger statistical mistakes and misinterpretations, against which we must guard.

A glance at my desktop explains much of the revolutionary change the advances in computing have brought. On it sits a modern computer workstation with 16 MB of memory, a 400 MB disk, a graphical user interface with multiple windows and multitasking, multi-media capabilities allowing me access to a compact disk reader and centrally archived statistical data files. On the compact disk reader, I can mount the latest product releases from the 1990 U.S. decennial census and, through my workstation and networked computing devices, I can reformat census data, analyze them in one or more of a half dozen statistical packages, and prepare publication quality tables, graphs, and maps. (Sadly, I note that the census files to which I have access have lost much of their useful detail because of the application by the U.S. Bureau of the Census of stringent disclosure-avoidance procedures.) The computing facilities on my desktop are not unique to my office or university. They can be found in most universities, businesses, and government statistical agencies, and if they are not present today in your office, they will be before too long.

Let us contrast this situation with that of 20 or 30 years ago. The results of the 1960 and 1970 U.S. decennial censuses were available primarily in paper form, and through a small number of specially funded and licensed state computer centers that provided special census tabulations to researchers and policy makers using elaborate mainframe computers, banks of tape drives, and relatively crude statistical programs that did little more than compute means and variances or prepare cross-classifications. Research and policy use of 1960 and 1970 census files required substantial resources and specialized computer centers. Startup costs were high and few could contemplate doing more than request a few limited tabulations. My personal workstation possesses far more memory, storage, and facilities than the entire specialized census computer center we had at the University of Minnesota in the early 1970s. In fact, today, *anyone* with a desktop personal computer and a modem can access and analyze masses of census and survey data from a variety of sources.

Researchers and other users of statistical data have long recognized the desirability for expanded access to statistical information collected by others. They see substantial value in such data, especially because of the relatively low cost of accessing them

as compared with the cost of collecting similar information *de novo*. Modern computing environments have provided expanded facility for the analysis of such data, but they also pose new dangers. The ready access of public data and the ability to search and link files raise new concerns about the privacy rights of individuals and the need to attend to issues of confidentiality. Even in the more restricted domain of statistical data bases, both public and private, there are renewed concerns about the confidentiality of individual records.

Statistical agencies and survey researchers have always been concerned about the need to preserve the confidentiality of respondents in order to ensure the quality of the statistical data that they provide, and these concerns have been heightened by the decline in response rates for censuses and surveys over the past two decades. There was an initial flurry of work on the topic disclosure and disclosure avoidance in the 1970s, e.g., see Rapaport and Sundgren (1975), Barabba and Kaplan (1975), and other papers at the 1975 International Statistical Institute meetings, as well as the symposium proceedings edited by Dalenius and Klevmarken (1976). Methodological advances then proceeded fitfully for a number of years, but attention to the issues has been heightened again over the past six years, as is evidenced by the growth in published papers on the topic, special issues of *Statistica Neerlandica* in 1992 and the *Journal of Official Statistics* in 1993, as well as a number of conferences devoted to the topic of privacy and confidentiality such as the one cosponsored by the International Statistical Institute and EUROSTAT in 1992.

The public is rightly concerned about personal information gathered by governments and by private researchers and what happens to it. A promise of confidentiality does have an effect on cooperation rates even though studies have shown that the respondents only vaguely understand the concept (see Panel on Privacy and Confidentiality as Factors in Survey Response 1979; Turner 1982; Singer 1983; Singer, Hippler, and Schwarz 1990). Even when survey responses are truly protected by law, respondents do not fully trust a confidentiality pledge. The public concern cuts both ways, however, and the principle of democratic accountability articulated in the new report by the Panel on Confidentiality and Data Access (Duncan, Jabine, and de Wolf 1993) also argues for the responsible dissemination of data to users.

In the remainder of this paper, I address several aspects of the issues of privacy and confidentiality as they pertain to statistical data bases. My perspective is primarily an American one, as I have spent the bulk of my professional career in the United States, but my comments reflect similar concerns and assessments that are taking place in Canada and in other countries around the world. In the next section, I begin with the broad research issues of the protection of research subjects and informed consent and then turn to the ethical and legal considerations that must be confronted to resolve the conflict between those laws and norms that dictate expanded access and the restrictions required to preserve privacy and confidentiality. One of the difficulties in addressing the conflict is interpreting in a technical fashion what is meant by the key notion confidentiality, which I do in Section 3. In Section 4, I give a brief overview of approaches for statistical data access that have been proposed that guarantee the preservation of confidentiality. In Section 5, I describe additional concerns that result from the use of regulatory data for statistical purposes and the potential use

of statistical data for regulation. Finally in Section 6, I briefly list a few additional issues.

2. Some Ethical and Legal Dimensions of Privacy and Confidentiality

Ethical issues associated with human experimentation and study, especially in medicine, have been of increasing concern over the past several decades and prescriptions for the protection of human subjects are typically enshrined in the doctrine of informed consent, and overseen in the U.S. by federally mandated institutional review boards who are required to approve research involving human subjects. Especially in university settings, these boards and committees are often asked to approve sample surveys and it is in this context that privacy and confidentiality issues typically are considered. It is common for university-based surveys to promise confidentiality of the data gathered to respondents, although there is rarely a legal statute that university researchers can rely upon to back up such guarantees. For example, U.S. courts have recently required scientists to surrender raw data without confidentiality protection when analyses of the data have been cited by parties in a lawsuit (see Marshall 1993). At the same time, the university and research community has also come to recognize the obligation it has to share research data (Fienberg, Martin, and Straf 1985), in part because of the scientific obligation to permit others to judge one's work and in part because of the societal value that accrues from data access. Thus the tension between confidentiality protection and expanded data access is a topic of heated debate (Panel on Scientific Research and the Conduct of Science 1992, pp. 47–49).

While the protection of confidentiality in university research settings has been discussed primarily in the context of individuals, in government agencies the concern extends quite rightly to establishments and other units of observation, for which the risks and consequences of disclosure are often substantial. Many of the disclosure avoidance rules described below were developed for censuses and surveys of establishments. Privacy and confidentiality concerns also extend to other forms of sampling not involving surveys as we commonly understand them. For example, Rathje and Murphy (1992) describe privacy and confidentiality concerns in studies of garbage gathered from households and landfills!

There has been a longstanding government interest and concern in the United States over the confidentiality of statistical data, especially as gathered in sample surveys and censuses. The U.S. Bureau of the Census operates under Title 13 of the U.S. Code, and, virtually from its inception in 1929, Title 13 had explicitly addressed the issue of the protection of the information gathered. The current language[2] prohibits the bureau from:

1. us[ing] the information furnished under the provisions of this title for any purpose other than the statistical purposes for which it is supplied; or
2. mak[ing] any publication whereby the data furnished by any particular establishment or individual under this title can be identified; or
3. permit[ting] anyone other than the sworn officers and employees... to examine the individual reports.

Most of the other U.S. statistical agencies have some form of confidentiality

[2] 13 USC 9.

protection as part of their legislative mandates, but few have as stringent language and approach as does the Bureau of the Census (Duncan, Jabine, and de Wolf 1993). Such legal guarantees of confidentiality are not only a reflection of the public concerns regarding disclosure but also of the agencies' desire for high quality data. As Vincent Barabba noted when he held the position of Director of the U.S. Bureau of the Census:

> The Bureau is zealous in pursuing the policy of confidentiality not just for legal and moral reasons, but also because of the simple fact that the data collection system ultimately depends on the goodwill and cooperation of people and companies. Should the public's confidence in the Bureau's pledge of confidentiality for their census returns erode, goodwill and cooperation will erode. (Barabba and Kaplan 1975)

At the same time government agencies have an obligation to report their data widely and thus they recognize the need for some balance between strict confidentiality (however it is to be interpreted) and the benefits derived from the release of statistical information.

Government agency pledges of confidentiality do not stand in isolation in the United States, and two major federal laws passed in the 1970s exemplify the tension between confidentiality and data access: The Privacy Act of 1974 and the Freedom of Information Act. The former prevents disclosure of records maintained on individuals while the latter prevents government agencies from refusing to provide public access to information (Norwood 1991; OFSPS 1978). But in weighing the balance between these two conflicting goals, some federal statistical agencies appear to have exhibited a clear bias towards withholding statistical information and microdata rather than providing it. This is the reason those advocating increased access to data have often pushed for explicit representation of disclosure risk and how an agency assesses it in the context of specific requests for data releases if the request is denied. To judge the balancing of disclosure risk and the benefits derived from the release of data the statistician requires a technical interpretation of words such as privacy, confidentiality, and disclosure, and we turn to this issue in the next section.

3. Some Formal Definitions and Technical Specifications

By individual privacy, we typically mean the freedom of the individual to decide how much of the self is to be revealed to others, when and to whom (Bulmer 1985). If information is shared with others, as in the setting of a survey or a census, then the notion of privacy extends to cover guarantees imparted to the individual when information is collected. Confidentiality reflects the desire of an individual to restrict access to personal information by others or the purposes to which it can be put. Privacy can be thought of as a state of the person, and thus the right to privacy is a personal one, although it might also be thought of as a property right. In contrast, confidentiality is a state of the information (Rieken 1983). When harm or dysfunction is the consequence of disclosure of confidential information on an individual, then the invasion of privacy becomes especially problematic (c.f., Barabba 1975).

In the U.S., the Privacy Act provides for disclosure without the consent of the individual to whom the information pertains only with the "advance adequate written assurance that the information will be used solely as a statistical research or reporting record, and the record is to be transferred in a form that is not individually identifiable" (Section 552a(b)(5)). In

interpreting this language, the U.S. Office of Management and Budget has indicated that this "means not only that the information disclosed or transferred must be stripped of individual identifiers but also that the identity of the individual cannot be reasonably deduced by anyone from tabulations or other presentations of the information (i.e., the identity of the individual cannot be determined or deduced by combining various statistical records or by reference to public records or other available sources of information)" (U.S. Office of Management and Budget 1975). We discuss the technical interpretation of such notions in a moment.

As we noted above, the concepts of privacy and confidentiality extend to establishments and organizations, although their rights to privacy do not have the same fundamental basis and are often subject to greater regulation by government. The boundary between data gathered for statistical purposes and data gathered for regulatory purposes is especially problematic and in a later section we discuss the problems at greater length.

A key notion in all of these concepts is that of disclosure, for which various authors have attempted to provide a precise definition. For example, Fellegi (1972) suggests that disclosure requires both the recognition of an individual member of a population included in a data release *and* learning something about that individual. In the context of sample surveys, the first part of the definition would mean that someone could actually identify a sample member on the basis of the data release without knowing a priori that the individual was a member of the sample. The second part means that the act of identifying someone as a sample member by a unique set of characteristics is not, in and of itself, a disclosure without there being the release of additional characteristics which are then identifiable. Fellegi goes on to discuss the notions of direct and residual disclosure.

Here, we advocate the adoption of a somewhat broader definition proposed originally by Dalenius (1977) and slightly reworded by Steinberg (1983):

If the release of certain statistical information makes it possible to determine a particular value relating to a known individual more accurately than is possible without access to that data, then a disclosure has taken place.

Duncan and Lambert (1989) describe this notion as *inferential disclosure* and contrast it with other notions of disclosure proposed in the literature. This definition is essentially a probabilistic one and is related to other probabilistic definitions such as those proposed by Cassel (1976) and Frank (1978).

We now formalize the Dalenius definition. Let S be the data released and E the information already known or available to a given user. Further, suppose that we are concerned about the disclosure of a characteristic or property of an individual. Let D be a characteristic, which may be a count or a magnitude measured by the survey or it may be some other characteristic and let D_k be the value of D assumed by the kth unit, O_k. For simplicity, suppose that $D_k = 1$ if O_k has a certain property, and $D_k = 0$ otherwise. Then a disclosure occurs if

$$\Pr(D_k = 1|S, E) > \Pr(D_k = 1|E). \qquad (1)$$

This definition extends to magnitudes, and in particular to the disclosure that an individual's value, D_k, lies within some interval. Because almost any data release provides some information about D_k, the total avoidance of disclosure is impossible. While confidentiality legislation is often written to imply zero disclosure risk, it is

clear that all forms of disclosure cannot be avoided as long as some information is actually released. Thus we are left with the approach of controlling or limiting disclosure. We also note that we can have a disclosure according to equation (1) for someone in the population who has not actually provided data. Thus a disclosure does not always produce a breach of confidentiality.

There is general agreement that, after specific identifiers, geographic information poses one of the greatest risks for disclosure. Steinberg (1983) suggests the simple example of a user being able to deduce information about a particular physician's income from a table that contains no clear identification of individuals but that shows a distribution of income, by occupation, for each city ward. For reasons such as this, statistical agencies have tended to develop rules on the suppression of detailed geographic information. For example, the Census Bureau typically does not provide geographic detail as part of microdata files when an area has fewer than 100,000 persons in the sample frame (Greenberg and Zayatz 1992). Thus, in public-use microdata tapes for the National Crime Survey (conducted by the Census Bureau for the Bureau of Justice Statistics) the absence of detailed local geographic information prevents the kinds of statistical analyses that would explore "ecological" correlates of crime victimization. Recent work at the Census Bureau attempts to address this analytical concern through the creation for microdata files of "contextual" variables that present reduced disclosure risk (Saalfeld, Zayatz, and Hoel 1992). There is also considerable agreement that longitudinal information, such as is available from the Survey of Income and Program Participation, makes for easy identification of individual microdata files, but sensible reporting rules for avoiding disclosure are harder to develop for longitudinal records.

We note the special role played in expression (1) by external information, E, e.g., information available to the user from a population register. Of special concern in this regard is the release of aggregate information in multiple and perhaps overlapping forms which, when combined, allow for disclosure in the probabilistic sense. Thus for a given release of statistical information, S, all prior releases of information from the same data base could be viewed as forming part of E (even though this appears to be precluded in the formal framework originally presented by Dalenius).

The Office of Federal Statistical Policy and Standards (OFSPS) (1978) gives a careful typology of instances of disclosure according to the Dalenius definition, including exact, approximate, probability based, and indirect disclosures, for macrodata (e.g., tabulations and other summary information) and for microdata. The paper distinguishes between external disclosure, e.g., to someone who is not a member of a particular cell in a tabulation, and internal disclosure to another member of that cell. For example, suppose that in a cross-classification of firms there is a cell with a count of 2. Then if firm A falls into that cell, the release of the cross-classification produces an internal disclosure to firm A of the information associated with the other firm in that cell.

The practical effect of a disclosure, in the technical sense described above, is a function of the magnitude of the disclosure, which is measured by the posterior probability $\Pr(D_k = 1|S, E)$, as well as the extent of disclosure measured by the difference

$$\Pr(D_k = 1|S, E) - \Pr(D_k = 1|E) = \text{posterior} - \text{prior}. \quad (2)$$

Duncan and Lambert (1986, 1989) pursue this probabilistic notion of disclosure by applying an uncertainty function (DeGroot 1962) to the probabilities, i.e., any nonnegative concave function $U(\bullet)$. They then focus on quantities such as *knowledge gain*,

$$U(\text{posterior}) - U(\text{prior}) \quad (3)$$

or *relative knowledge gain*

$$(U(\text{posterior}) - U(\text{prior}))/U(\text{prior}). \quad (4)$$

Factors affecting the risk of disclosure, in addition to the prior information E, include the choice of variables to be reported, the age of the data files, and the aspects of nonsampling error and their effects. The extent of disclosure might be so slight as to be essentially undetectable. Furthermore, agencies responsible for data often do not have full knowledge about E, the external information available to others. The risk of disclosure in a release S for population data, as in a census, is clearly greater than for exactly the same kinds of data releases for sample data. Indeed, the real aim of an organization addressing the issue of confidentiality is the exercise of disclosure control and the acceptability of disclosure risk associated with various kinds of data in different situations. Duncan and Lambert (1986) illustrate how this probabilistic approach can be applied to provide a justification for various ad hoc rules proposed or actually used to limit disclosure. A number of proposals in the literature can be recast in terms of the Dalenius definition and the probabilistic or uncertainty difference measures. For example, Frank (1978) and Greenberg and Zayatz (1992) describe an approach to measuring relative disclosure risk in terms of the entropy. But Shannon's entropy is a member of the class of uncertainty functions to be applied to posterior probabilities in the Duncan and Lambert framework.

OFSPS (1978) provides an excellent discussion of the issues and trade-offs associated with assessing disclosure risk, and it describes a variety of statistical agency approaches to and policies on disclosure avoidance in the 1970s. Most of these are ad hoc in nature and are intentional, conservative attempts to deal with the kinds of issues associated with the more formal framework described in this section. Except for the work of Duncan and Lambert (1986), we have yet to see a systematic interpretation of an agency's disclosure rules in terms of a technical probabilistic framework such as that set out here, or an agency attempt to translate a given set of rules into a set of probabilistic statements that can be interpreted and understood by survey respondents and by those requesting access to survey data.

Lambert (1993) argues that one can make an assessment of the ad hoc rules currently in use *only when* we have a working model for the behavior of the intruder. This is the approach adopted in Fienberg and Makov (1993), who use a formal Bayesian framework consistent with that introduced in this section.

4. Statistical Disclosure-Avoidance Options for Microdata

4.1. *Identification and uniqueness*

For many, the issue of disclosure is linked to the possibility of identification of individuals through certain *key variables*, called *identity disclosure* by Duncan and Lambert (1989), and the resulting release of sensitive information on those individuals. For example, Bethlehem, Keller, and Pannekoek (1990) argue that a disseminated microdata set should be constructed

so that it is impossible for others to link records to individuals correctly by using identifying information in the data set and prior knowledge. This leads them to focus on the "uniqueness" of individuals according to the key variables that might be used for identification (see also the early work of Olsson and Block (1976) as well as Dalenius (1986)). An individual is unique in a population if he/she is the only one in the population with a particular set of values on the key variables. Similarly, an individual is unique in a sample if he/she is the only one in the sample with a particular set of values on the key variables. Uniqueness in the population implies uniqueness in the sample but not conversely. A microdata set composed of a sample from a population thus may contain "uniques" who are not unique in the population. Bethlehem et al. (1990) consider the problem of estimating uniqueness in the population using sample data and a superpopulation model. They then attempt to relate their estimator to disclosure risk for microdata.

Avoidance of identity disclosure can be placed into a probabilistic framework, e.g., see Duncan and Lambert (1989), but it is not equivalent to Dalenius's inferential disclosure approach. Several authors have recently explored different aspects of identity disclosure in microdata files from different European countries. Paass (1988) reports on disclosure experiments using German government data (and large numbers of variables) with different search strategies and finds high levels of identity disclosure, even with somewhat noisy data (but see the comments on his calculations in Duncan and Lambert 1989). Marsh et al. (1991) consider the probability of disclosure as a product of four components and conclude that the disclosure risk for a planned microdata sample from the British census is on the order of 1 in 4 million. Skinner (1992) explores the somewhat weaker notion of prediction disclosure and its relationship to identity disclosure for microdata samples. Finally, Biggeri and Zannella (1991) report on preliminary results from simulation studies using Italian microdata files. It is difficult to draw any systematic conclusions from these studies, except to note that, the larger the number of variables on the microdata files, the greater the risk of disclosure.

Despite the claims made by many authors regarding the dangers of disclosure, there are in fact few documented cases of serious disclosure of sensitive information by a modern statistical agency. As Sundgren (1993, p. 512) notes, "If such a leak had actually occurred, one could be reasonably sure that some alert newspaper would enthusiastically have scandalized the failing statistical office publicly." This suggests that most agency rules may have overemphasized disclosure risk and have been less concerned with access to data and the utility of data released. The disclosure experiments reported by Blien, Wirth, and Müller (1992), in which scientific intruders had enormous difficulty in identifying uniques in anonymized microdata files largely due to errors in the data, lend support to this view.

4.2. General approaches

No matter which definition of disclosure one uses, the release of microdata may pose serious risks, depending on the numbers of variables involved and their information content. The four most frequently proposed solutions to the preservation of data confidentiality in such circumstances (aside from no release) are:

1. *Remote access.* In this solution, statistical analyses are submitted by researchers over a computer network

to be run in a central location and checked for possible violations of confidentiality. The oft-cited example is the Luxembourg Income Study. See the discussion in Duncan and Pearson (1991) and Smith (1991).

2. *Special sworn employees.* The U.S. Census Bureau has a tradition of making individuals special sworn employees and thus according them on-site access to selected confidential data in a manner similar to regular census employees. All data released as a result of such statistical activities must meet the usual agency restrictions, and the special employees cannot even take derivative analytical files that do not meet release criteria. In addition, the bureau must vouch that special sworn employees are working to serve the bureau's needs.
3. *Licensing of researchers.* Slightly less restrictive is the licensing of researchers to use sensitive data files for specific purposes and specified periods of time, under conditions that include penalties for improper use. This approach has been tried with the National Longitudinal Survey at Ohio State University and the Panel Survey of Income Dynamics at the University of Michigan. The National Center for Education Statistics has begun to use this approach as well. For further details, see Duncan and Pearson (1991).
4. *Statistical models.* This approach releases either an "unidentifiable" minimal-disclosure summary of the data, such as a variance–covariance matrix or a multiple cross-classification, or a transformed version of original microdata (with identifiers removed). We describe it in detail below.

4.3. *Data masking for microdata*

Duncan and Pearson (1991) give an excellent description of approaches to the masking of microdata. Suppose that $\mathbf{X}$ is an n by p matrix representing the microdata for n individuals or cases on p variables or attributes. Then matrix masking of the microdata file $\mathbf{X}$ provides the user with the transformed file

$$\mathbf{M} = \mathbf{AXB} + \mathbf{C} \tag{5}$$

in lieu of $\mathbf{X}$. The matrix $\mathbf{A}$ transforms cases, $\mathbf{B}$ transforms variables, and $\mathbf{C}$ blurs the entries of $\mathbf{X}$ or more generally $\mathbf{AXB}$. The use of $\mathbf{M}$ in lieu of $\mathbf{X}$ includes several well-known approaches as special cases:

1. Release a subset or sample of the data (delete rows of $\mathbf{X}$).
2. Include simulated data (add rows to $\mathbf{X}$).
3. Add random perturbations to $\mathbf{X}$.
4. Exclude selected attributes (delete columns of $\mathbf{X}$).
5. Release the variance–covariance matrix (choose $\mathbf{A} = \mathbf{X}^{\mathrm{T}}$).

Examples of transformations to $\mathbf{X}$ that are not of the form $\mathbf{M}$ include swapping (exchanging rows for a subset of the columns of $\mathbf{X}$) and the coarsening, grouping or truncation of attributes.

Clearly the use of $\mathbf{M}$ needs some information about ($\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$), but the release of full information is not allowed. Determining what information can be released for a given choice of ($\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$) and the choice of $\mathbf{M}$ itself are both active areas of research. There are also issues regarding the effect of linkage of files on choices of ($\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$). For example, suppose that we wish to release information about two separate data bases as well as a merged or linked version of them using matrix masking. Just what information about the three different choices of ($\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$) can be released? All such questions must be addressed in the

context of the probabilistic definition of disclosure described in the previous section.

Duncan and Lambert (1989) provide an excellent description of how their disclosure limitation approach can be applied to the problem of releasing a transformation of microdata, but they do not directly address either the choice of the transformation, (**A**, **B**, **C**). or the information on it to be released. Fuller (1993) gives a detailed partially-Bayesian treatment of the properties of adding random normal perturbations (where **A** and **B** are identity matrices) to a multivariate normal vector, and he describes a measurement error approach to developing such additive masks for continuous or for categorical data due to Sullivan (1989). Little (1993) adopts a likelihood approach to the masking problem and exploits recent theory on treatment assignment and missing data. His approach makes explicit the role of the masking selection mechanism in the likelihood function, something which is also relevant for a Bayesian analysis. Little considers several special cases of matrix masking, as well as masking by coarsening. Rubin (1993) proposes a similar approach through the Bayesian method of multiple imputation.

The use of matrix masking techniques is widespread in U.S. statistical agencies, but critics claim that it provides only a partial answer to disclosure avoidance as it assumes that (i) all relevant data for statistical analyses do not come in the form of $n \times p$ matrices, and (ii) the desired forms of statistical analyses can be specified in advance, thereby allowing the determination of a suitable transformation and the information on it to be released (e.g., see Smith 1991). In fact, most data can be transformed into matrix form which is a requisite for masking. Consider, for example, longitudinal data of n individuals for p variables at q time periods. The data form an $n \times p \times q$ three-dimensional matrix that can be transformed into an $nq \times pq$ block-diagonal matrix, with many "structural zeros" but which does not lose any of the information in the original three-dimensional matrix. The problem is that special masks need to be designed for such arrays in order to facilitate sensible longitudinal analyses. Similarly, hierarchical data files can also be given flat file or square matrix representations. To our knowledge, however, limited effort has been expended on the technical representation of disclosure issues associated with the hierarchical longitudinally-linked data files such as those generated from the Survey of Income and Program Participation and from rotating panel surveys such as the Current Population Survey and the National Crime Survey.

How well do masking methods deal with the trade-offs between confidentiality and access? There appears to be a general presumption that, if **X** takes a particular form or a specific class of analyses is planned in advance, then a mask can be devised that both protects confidentiality and does not impede the analyses very much. On the other hand, most statisticians agree that masks designed as all-purpose protection devises are likely to get in the way of many standard approaches to data analysis, especially ones that emphasize robustness, outliers, and data leverage.

4.4. Cell suppression for cross-classifications

A special case involving the deletion of rows, *cell suppression*, is worthy of additional discussion. Suppose we are interested in summarizing a set of data in the form of a cross-classification of counts or nonnegative aggregates. Deleting or suppressing a cell value is equivalent to the

deletion of those rows of **X** for which the entries in columns corresponding to the cross-classifying variables assume the values that specify the cell in question. Cell suppression is widely used for data on establishments because counts of "1" may uniquely identify a firm in a given industrial sector. (See the discussion of uniqueness and identity disclosure above.)

Current practice at the U.S. Census Bureau is to suppress any cell where n or fewer respondents make up k or more percent of that cell's value. Such cells are referred to as *primary suppressions*. The bureau keeps the values of n and k as well as the method used for their selection confidential (Zayatz 1992), although such a practice may well be excessive when viewed from the perspective espoused in this paper. One way to view the choice of the values of n and k is via the Duncan–Lambert disclosure limitation framework.

Because reported cross-classifications usually include the corresponding marginal totals, suppressing a single cell produces multiple masks of the same matrix and, taken together, these masks do not disguise the data – the value of a deleted cell in a two-way array can be retrieved from the other entries in the same row or column combined with the corresponding marginal total. Thus methods for cell suppression in cross-classifications also choose other cell values for suppression; these are often referred to as *complementary suppressions*. Determining "desirable" patterns of complementary suppressions is an active area of research, especially for multi-way cross-classifications (e.g., see Zayatz 1992; Sullivan and Rowe 1992). Multiple primary suppressions combined with complementary suppressions may also produce suppressed values in the marginal totals. Little (1993) considers an application of his likelihood approach to the suppression problem and also discusses imputation as an alternative to suppression. The Duncan–Lambert disclosure limitation approach can also be used to study the choice of appropriate complementary suppressions.

5. Statistical Versus Administrative Uses of Data

Government agencies and researchers often turn to data originally collected for administrative or regulatory purposes in order to carry out statistical analyses. Thus the Social Security Administration (SSA) or the Internal Revenue Service (IRS) will often use agency data collected in the administration of their day-to-day activities to investigate larger issues, or at least ones not directly related to the individuals or establishments that provided the original information. For example, IRS conducts periodic samples of tax returns and then carries out full audits on them in order to determine operational rules on when to conduct audits of other returns. These data have been used by various researchers to study issues related to deterrence (e.g., see Clotfelter 1983; Klepper, Mazur, and Nagin 1991; Klepper and Nagin 1989; Poterba 1987). Statistically derived files from administrative records may be matched and linked with other statistical data bases to create new files with more individual information than that possessed by the original administrative agency. Such statistical data bases create additional problems for disclosure prevention since different agencies have different disclosure policies and rules.

When statistical data bases are derived from administrative or regulatory data there is the added issue of reverse access. Suppose an agency matches tax records from the IRS with other information in a separate statistical file, thus appending

variables to the original files for at least some individuals. The resulting files may well have new administrative or regulatory value and the information contained therein may, if provided to the IRS, adversely affect the individuals' tax liabilities. Special protection is required to preserve the confidentiality of such files.

As a general rule, statisticians have argued that, while administrative data can be used for statistical purposes, statistical data must not be used for regulatory purposes. This has been described as the principle of functional separation (Duncan, Jabine, and de Wolf 1993). A recent example involving establishment data from the Energy Information Agency (EIA) illustrates the problems raised by the violation of this principle and the fact that the statistician's perspective on the confidentiality of statistical data is not always shared by others in government. For a related description see Kirkendall (1992).

EIA collects information for statistical purposes from individual oil companies on prices, costs, capacity, and output, under the authority of the Federal Energy Administration Act[3]. Some of the major oil companies are known to be included in EIA surveys with certainty while others (smaller independent companies) are included with probabilities less than one. EIA has a disclosure policy[4] which restricts access to these data by other federal agencies unless the individual companies consent, a court has ordered the disclosure, or the President has so directed.

A few years ago, the Antitrust Division of the Department of Justice, in its investigation of gasoline and heating oil pricing, requested access to individually identifiable company data collected by EIA, citing as its authority another section of the Federal Energy Administration Act.[5] Relying upon its confidentiality policy, EIA refused the request. The Department of Justice Office of Legal Counsel, when asked for its opinion, responded that EIA was required to provide the requested information. EIA's concern related not only to its obligations to the individual oil companies under its original pledge of restricted confidentiality but also to the level of cooperation it was likely to achieve in future surveys. Others in and out of the federal government expressed grave concerns regarding the precedent that would be set by the surrender of individual statistical data for regulatory purposes.

Before this dispute could be settled, the Department of Justice made a different request for access to data for enforcement purposes linked to an investigation associated with the Gulf War. EIA did not respond with the data and the Department of Justice found an alternative source for the information in this second request. Finally, with the change in administration in 1993, the original request was dropped.

While EIA did not surrender the data it believed were protected under its pledge of confidentiality, damage was done to its credibility, in terms of the public perception of the confidentiality of EIA data. Furthermore the legal status of EIA's confidentiality guarantees remains in doubt and a cause for heightened concern amongst those working with statistical data bases, both inside and outside of U.S. statistical agencies (see the related discussion and recommendations in Duncan, Jabine, and de Wolf 1993).

6. Related Issues

Computers are not the only threat to the confidentiality of survey and other

[3] 18 USC 1905.

[4] 45 Federal Register 59, 812 (September 10, 1980).

[5] 15 USC 771(f)(1).

research data. The courts pose an additional threat. When a researcher conducts a study that is subsequently introduced (possibly by someone unrelated to the researcher) as evidence by one party in a legal proceeding, courts have ruled that the original researcher must provide the data for examination by the opposing party (e.g., Barinaga 1992). Many courts in the United States have recognized the need to preserve confidentiality for data produced in such circumstances and have authorized the removal of identifiers (but see examples of exceptions described by Marshall 1993). Unfortunately, issues such as those discussed in Sections 3 and 4 have not been prominent in such cases. The release of epidemiological information from retrospective studies involving rare diseases linked to specific products poses special problems since the names and other information about affected individuals may already be known. This was the situation in a celebrated legal battle over access to a study linking aspirin to the occurrence of Reyes' Syndrome in the 1980s (e.g., see Fienberg 1994).

A topic that continues to generate considerable debate and even strong emotion in the United States is the sharing of confidential data among statistical agencies. Unlike the problem of statistical versus regulatory uses of data discussed in the preceding section, what is at issue here is the sharing of data where (a) both agencies engage in only statistical activities and (b) the data shared would be covered by a legally-based guarantee of confidentiality in the recipient agency. For years, the Census Bureau resisted requests for such sharing, citing the language of Title 13 as forbidding it. About 15 years ago, the Carter Administration's Statistical Reorganization Project proposed legislation that would address this issue through the creation of "protected statistical centers." The legislation was never enacted, however, and the difficulty of sharing confidential data among agencies still exists.

Establishment data raise special concerns regarding disclosure, as we have noted in previous sections, and government statistical agencies in the United States have been especially cautious in releasing information from establishment surveys. Because of the language in Title 13, the Census Bureau had long resisted sharing its Standard Statistical Establishment List of U.S. businesses with other statistical agencies (e.g., see the discussion in Alexander 1983), although there is now an agreement between the Census Bureau and the Bureau of Labor Statistics on sharing information about Standard Industrial Codes (SIC).

Finally, we note that the Census Bureau has recently suggested that the confidentiality protection provided by Title 13 extends to all address lists of housing units that it prepares, even if that information is not necessarily gathered from specific individuals or under a pledge of confidentiality. This is likely to be a topic of substantial debate in the next several years, both because of interest in the use of administrative records for census purposes and because of the bureau's own attempts to market its Topologically Integrated Encoding and Referencing (TIGER) system, developed originally for the 1990 decennial census. TIGER includes as complete an address listing for households in the United States as the bureau possesses and different pieces of information in TIGER come from diverse sources, including information collected in connection with the 1990 decennial census, but not necessarily under representations of confidentiality made by the bureau under Title 13. For example,

see the discussion of this issue in Panel on Census Requirements in the Year 2000 and Beyond (1993).

7. Summary

With the growth of computer-based government records and the collection of statistical data for research, and the new era of computing and telecommunications that has emerged in the past decade, the demand for greater access to data has increased dramatically. So too has the risk that someone who gains access to a data set will either intentionally or inadvertently identify individuals and information about them. Statistical agencies and survey researchers have always been concerned about the need to preserve the confidentiality of respondents in order to ensure the quality of the data provided, and these concerns have been heightened by the decline in response rates for censuses and surveys over the past two decades.

In this paper I have attempted to discuss the conflict between the perspective of access to data, on the one hand, and demands for confidentiality on the other. In doing so, I have adopted a technical definition of disclosure and suggested that it be used to formally assess the trade-offs between access and confidentiality, and to understand the effect of disclosure-avoidance methods such as matrix masking and cell suppression. I do not have a prescription on assessing such trade-offs, but I believe that a fully Bayesian perspective, viewing disclosure from the perspective of an intruder, holds the most promise. Further, I believe that we would all benefit from detailed case studies, ones where there is far greater detail than is typically afforded by a journal article or a conference proceedings paper, since no single approach will suffice for all organizations and agencies or even all data sets within an agency.

8. References

Alexander, L. (1983). Proposed Legislation to Improve Statistical and Research Access to Federal Records. Chapter 15 in Solutions to Ethical and Legal Problems in Social Research, R.F. Boruch and J.S. Cecil, eds., New York: Academic Press, 273–292.

Barabba, V.P. (1975). The Right of Privacy and the Need to Know. In The Census Bureau: A Numerator and a Denominator for Measuring Change. U.S. Bureau of the Census Technical Paper 37, Washington, DC: U.S. Department of Commerce, 23–29.

Barabba, V.P. and Kaplan, D.L. (1975). U.S. Census Bureau Statistical Techniques to Prevent Disclosure – The Right of Privacy vs. the Need to Know. Paper presented at the 40th Session of the International Statistical Institute, Warsaw, Poland.

Barinaga, M. (1992). News and Comment: Who Controls a Researcher's Files? Science, 256, 1620–1621.

Bethlehem, J.C., Keller, W.J., and Pannekoek, J. (1990). Disclosure Control of Microdata. Journal of the American Statistical Association, 85, 38–45.

Biggeri, L. and Zannella, F. (1991). Release of Microdata and Statistical Disclosure Control in the New National Statistical System of Italy: Main Problems, Some Technical Solutions, Experiments. Bulletin of the International Statistical Institute, 48th Session, Cairo, Egypt.

Blien, U., Wirth, H., and Müller, M. (1992). Disclosure Risk from Microdata Stemming from Official Statistics. Statistica Neerlandica, 46, 69–82.

Bulmer, M. (1985). Ethics in Social Research. In The Social Science Encyclopedia, A. Kuper and J. Kuper, eds., Routledge & Kegan Paul, London, 265–267.

Cassel, C.M. (1976). Probability Based Disclosures. In Personal Integrity and the Need for Data in the Social Sciences, T. Dalenius and A. Klevmarken, eds., Stockholm: Swedish Council for the Social Sciences, 189–193.

Clotfelter, C.T. (1983). Tax Evasion and Tax Rates: An Analysis of Individual Returns. Review of Economics and Statistics, 65, 363–373.

Dalenius, T. and Klevmarken, A., eds. (1976). Personal Integrity and the Need for Data in the Social Sciences. Stockholm: Swedish Council for the Social Sciences.

Dalenius, T. (1977). Towards a Methodology for Statistical Disclosure Control. Statistisk tidskrift, 5, 429–444.

Dalenius, T. (1986). Identifying Anonymous Census Records. Journal of Official Statistics, 2, 329–336.

DeGroot, M.H. (1962). Uncertainty, Information, and Sequential Experiments. Annals of Mathematical Statistics, 33, 404–419.

Duncan, G.T., Jabine, T.B., and de Wolf, V.A., eds. (1993). Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics. Panel on Confidentiality and Data Access, Committee on National Statistics. Washington, DC: National Academy Press.

Duncan G.T. and Lambert, D. (1986). Disclosure-Limited Data Dissemination (with discussion). Journal of the American Statistical Association, 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.

Duncan, G.T. and Pearson, R.B. (1991). Enhancing Access to Microdata While Protecting Confidentiality: Prospects for the Future (with discussion). Statistical Science, 6, 219–239.

Fellegi, I.P. (1972). On the Question of Statistical Confidentiality. Journal of the American Statistical Association, 67, 7–18.

Fienberg, S.E. (1994). Sharing Statistical Data in the Biomedical and Health Sciences: Ethical, Institutional, Legal, and Professional Dimensions. Annual Review of Public Health, 15, 1–18.

Fienberg, S.E. and Makov, U.E. (1993). A Bayesian Approach to Data Disclosure. Bulletin of the International Statistical Institute, 49th Session, Florence, Italy.

Fienberg, S.E., Martin, M.E., and Straf, M.L., eds. (1985). Sharing Research Data. Committee on National Statistics. Washington, DC: National Academy Press.

Frank, O. (1978). An Application of Information Theory to the Problem of Statistical Disclosure. Journal of Statistical Planning and Inference, 2, 143–152.

Fuller, W. (1993). Masking Procedures for Microdata Disclosure. Journal of Official Statistics, 9, 383–406.

Greenberg, B.V. and Zayatz, L.V. (1992). Strategies for Measuring Risk in Public Use Microdata Files. Statistica Neerlandica, 46, 33–48.

Kirkendall, N.J. (1992). Data Access or Protection? An example of an Interagency Disagreement. Proceedings of the Social Statistics Section, American Statistical Association, 17–21.

Klepper, S., Mazur, M., and Nagin, D. (1991). Expert Intermediaries and Legal Compliance: The Case of Tax Preparers. Journal of Law and Economics, 34, 205–229.

Klepper, S. and Nagin, D. (1989). The Anatomy of Tax Evasion. Journal of

Law, Economics, and Organization, 5, 1–24.

Lambert, D. (1993). Measures of Disclosure Risk and Harm. Journal of Official Statistics, 9, 313–331.

Little, R.J.A. (1993). Statistical Analysis of Masked Data. Journal of Official Statistics, 9, 407–426.

Marsh, C., Skinner, C., Arber, S., Penhale, B., Openshaw, S., Hobcraft, J., Lievesley, D., and Walford, N. (1991). The Case for Samples of Anonymized Records from the 1991 Census. Journal of the Royal Statistical Society, Ser. A, 154, 305–340.

Marshall, E. (1993). News and Comment: Court Orders "Sharing" of Data. Science, 261, 284–286.

Norwood, J.L. (1991). Comment on a Paper by Duncan and Pearson. Statistical Science, 6, 236–237.

Office of Federal Statistical Policy and Standards (OFSPS) (1978). Report on Statistical Disclosure and Disclosure-Avoidance Techniques. Statistical Policy Working Paper 2. U.S. Department of Commerce. Washington, DC: U.S. Government Printing Office.

Olsson, L. and Block, H. (1976). Backwards Identification. Statistisk tidskrift, 14, 135–144, 168.

Panel on Census Requirements in the Year 2000 and Beyond (1993). Planning the Decennial Census. Interim Report. Committee on National Statistics Washington, DC: National Academy Press.

Panel on Privacy and Confidentiality as Factors in Survey Response (1979). Privacy and Confidentiality as Factors in Survey Response. Committee on National Statistics. Washington, DC: National Academy of Sciences.

Panel of Scientific Research and the Conduct of Science (1992). Responsible Science. Ensuring the Integrity of the Research Process. Volume I. Committee on Science, Engineering, and Public Policy. Washington, DC: National Academy Press.

Paass, G. (1988). Disclosure Risk and Disclosure Avoidance for Microdata. Journal of Business and Economic Statistics, 6, 487–500.

Poterba, J.M. (1987). Tax Evasion and Capital Gains Taxation. American Economic Review, 77, 234–239.

Rapaport, E. and Sundgren, B. (1975). Output Protection in Statistical Data Bases. Bulletin of the International Statistical Institute, 40th Session, Warsaw, Poland.

Rathje, W. and Murphy, C. (1992). Rubbish! The Archeology of Garbage. New York: HarperCollins.

Rothfeder, J. (1992). Privacy for Sale. How Computerization Has Made Everyone's Life an Open Secret. New York: Simon & Schuster.

Rieken, H.W. (1983). Solutions to Ethical and Legal Problems in Social Research: An Overview. Chapter 1 in Solutions to Ethical and Legal Problems in Social Research, R.F. Boruch and J.S. Cecil, eds. New York: Academic Press, 1–9.

Rubin, D.B. (1993). Satisfying Confidentiality Constraints Through the Use of Synthetic Multiply-Imputed Microdata. Journal of Official Statistics, 9, 461–468.

Saalfeld, A., Zayatz, L.V., and Hoel, E. (1992). Contextual Variables via Geographic Sorting: A Moving Averages Approach. Proceedings of the Section on Survey Research Methods, American Statistical Association, 691–696.

Singer, E. (1983). Informed Consent Procedures in Surveys: Some Reasons for Minimal Effects on Response. Chapter 10 in Solutions to Ethical and Legal Problems in Social Research, R.F. Boruch and J.S. Cecil, eds. New York: Academic Press, 183–211.

Singer, E., Hippler, H.-J., and Schwarz, N. (1990). The Effects of Confidentiality Assurances on Response. Paper presented at the International Conference on Measurement Errors in Surveys, November 1990, Tucson, Arizona, U.S.A.

Skinner, C. (1992). On Identification Disclosure and Prediction Disclosure for Microdata. Statistica Neerlandica, 46, 21–32.

Smith, J.P. (1991). Data Confidentiality: A Researcher's Perspective. Panel on Privacy and Confidentiality. Paper presented at the Annual Meeting of the American Statistical Association, Anaheim, CA.

Steinberg, J. (1983). Social Research Use of Archival Records: Procedural Solutions to Privacy Problems. Chapter 13 in Solutions to Ethical and Legal Problems in Social Research, R.F. Boruch and J.S. Cecil, eds., New York: Academic Press, 249–261.

Sullivan, G. (1989). The Use of Added Error to Avoid Disclosure in Microdata Releases. Unpublished Ph.D. dissertation, Department of Statistics, Iowa State University, Ames, Iowa.

Sullivan, C.M. and Rowe, E.G. (1992). A Data Structure and Integer Programming Technique to Facilitate Cell Suppression Strategies. Proceedings of the Section on Survey Research Methods, American Statistical Association, 685–690.

Sundgren, B. (1993). The Problem of Data Security and Confidentiality in Statistics Production: Where Do We Stand after 25 Years of Research. Journal of Official Statistics, 9, 511–517.

Turner, A.G. (1982). What Subjects of Survey Research Believe about Confidentiality. Chapter 7 in The Ethics of Social Research. Surveys and Experiments, J.E. Sieber, ed., New York: Springer-Verlag, 151–165.

U.S. Office of Management and Budget (1975). Privacy Act Implementation. Guidelines and Responsibilities. Federal Register, Part III, Vol. 40, No. 132, July, Washington, DC: Government Printing Office, 28948–28978.

Zayatz, L.V. (1992). Linear Programming Methodology Used for Disclosure Avoidance Purposes at the Census Bureau. Proceedings of the Section on Survey Research Methods, American Statistical Association, 674–684.

Received November 1992
Revised September 1993