Statistical Policy Working Paper 2

Report on Statistical Disclosure and Disclosure-Avoidance Techniques

1978
U.S. DEPARTMENT OF COMMERCE
Office of Federal Statistical Policy and Standards

Statistical Working Papers are a series of technical documents prepared under the auspices of the Office of Federal Statistical Policy and Standards. These documents are the product of working groups or task forces, as noted in the Preface to each report.

These Statistical Working Papers are published for the purpose of encouraging further discussion of the technical issues and to stimulate policy actions which flow from the technical findings. Readers of Statistical Working Papers are encouraged to communicate directly with the Office of Federal Statistical Policy and Standards with additional views, suggestions, or technical concerns.

Office of
Federal Statistical
Policy and Standards

Joseph W. Duncan
Director

Statistical Policy Working Paper 2

Report on Statistical Disclosure and Disclosure-Avoidance Techniques

Prepared by
Subcommittee on Disclosure-Avoidance Techniques
Federal Committee on Statistical Methodology

U.S. DEPARTMENT OF COMMERCE
Juanita M. Kreps, Secretary
Courtenay M. Slater, Chief Economist

Office of Federal Statistical Policy and Standards
Joseph W. Duncan, Director

Issued: May 1978

Office of Federal Statistical Policy and Standards

Joseph W. Duncan, *Director*

George E. Hall, *Deputy Director,* Social Statistics
Gaylord E. Worden, *Deputy Director,* Economic Statistics
Maria E. Gonzalez, *Chairperson,* Federal Committee on Statistical Methodology

Preface

This working paper was prepared by the members of the Subcommittee on Disclosure-Avoidance Techniques, Federal Committee on Statistical Methodology. The Subcommittee was chaired by John A. Michael, National Center for Education Statistics, Department of Health, Education, and Welfare. The members of the Subcommittee are the authors of this report and their names are listed below. This report is intended to help managerial and technical staff of Federal agencies which publish or otherwise release data, on methodologies to achieve appropriate disclosure-avoidance practices. Data released both in tabulations and in the form of microdata are discussed in this report. The Office of Federal Statistical Policy and Standards hopes to organize, with the help of Subcommittee members, seminars with Federal employees to disseminate the findings of the report. In addition, the report may serve as a basis for discussions between Federal data producers and data users.

Members of the Subcommittee on Disclosure-Avoidance Techniques

John A. Michael, *Chairperson*
National Center for Education Statistics (HEW)

Richard A. Bell
Social Security Administration (HEW)

Robert H. Mugge
National Center for Health Statistics (HEW)

Mervyn R. Stuckey
Statistical Reporting Service (USDA)

Thomas B. Jabine *
Social Security Administration (HEW)

William J. Smith, Jr.
Internal Revenue Service (Treasury)

Paul T. Zeisset
Bureau of the Census (Commerce)

Ex Officio

Maria Elena Gonzalez, *Chairperson* *
Federal Committee on Statistical Methodology,
Office of Federal Statistical Policy and Standards
(Commerce)

Tore E. Dalenius
Brown University and University of Stockholm

* Member, Federal Committee on Statistical Methodology

Acknowledgements

The body of this report represents the collective effort of the Subcommittee on Disclosure-Avoidance Techniques.

The Subcommittee began by developing the outline for this report, after which writing assignments were apportioned among members. Manuscript was usually subjected to several rounds of review before its acceptance. The major contributors to the respective chapters appear below:

Chapter	*Major Contributor(s)*
I	Michael
II	Jabine and Dalenius
III	Bell, Mugge, and Dalenius
IV	Zeisset
V	Michael and Zeisset
VI	Jabine

Appendix	
A	The respective agencies
B	Stuckey
C	Lawrence H. Cox, Bureau of the Census
D	Bell

Throughout the development of the report, Thomas Jabine enlightened Subcommittee members on the complexities of the subject and Maria Gonzalez provided encouragement and goal directedness. Members of the Federal Committee on Statistical Methodology and the Office of Federal Statistical Policy and Standards, Department of Commerce (formerly the Statistical Policy Division of OMB) reviewed and commented upon our work. Manuscript was prepared with the good-natured assistance of the management and secretaries of the various statistical agencies. Deserving special commendation is Joyce Peoples of the Social Security Administration who effectively managed the arduous task of preparing and assembling several drafts of this manuscript.

Members of the Federal Committee on Statistical Methodology

Barbara A. Bailar
Bureau of the Census (Commerce)

Norman D. Beller
Statistical Reporting Service (USDA)

Barbara A. Boyes
Bureau of Labor Statistics (Labor)

Edwin J. Coleman
Bureau of Economic Analysis (Commerce)

John E. Cremeans
Bureau of Economic Analysis (Commerce)

Marie D. Eldridge
National Center for Education Statistics (HEW)

Fred J. Frishman
International Revenue Service (Treasury)

Maria E. Gonzalez, *Chairperson*
Office of Federal Statistical Policy and Standards (Commerce)

Thomas B. Jabine
Social Security Administration (HEW)

Charles D. Jones
Bureau of the Census (Commerce)

Alfred D. McKeon
Bureau of Labor Statistics (Labor)

Harold Nisselson
Bureau of the Census (Commerce)

Monroe G. Sirken
National Center for Health Statistics (HEW)

Wray Smith
Office of the Assistant Secretary for Planning and Evaluation (HEW)

Editorial Note

The opinions expressed in this report reflect the collective judgment of the Subcommittee and do not necessarily reflect the opinion of the Federal Committee or the Office of Federal Statistical Policy and Standards.

Table of Contents

CHAPTER I—INTRODUCTION

CHAPTER II—DEFINING STATISTICAL DISCLOSURE

CHAPTER III—DISCLOSURE IN THE RELEASE OF TABULATIONS (SUMMARY DATA) FOR PUBLIC USE

CHAPTER IV—DISCLOSURE IN MICRODATA

CHAPTER V—THE QUESTION OF BALANCE: PROTECTION OF INDIVIDUALS VS. PUBLIC NEEDS FOR INFORMATION

CHAPTER VI—FINDINGS AND RECOMMENDATIONS

Page

APPENDICES

Introduction

A. Scope of Study and Organization of Report

This report is about techniques for avoiding disclosure of confidential information about individuals (natural and legal persons) in connection with the release of statistical tabulations and microdata files (computerized records pertaining to individual statistical units). The report culminates more than a year's study of potentials for statistical disclosure—i.e. disclosure of confidential information about identifiable (but not identified) units in tabulations and microdata files. Many Federal agencies which release tabulations or microdata files for statistical purposes have statutes, regulations, or policy requirements that releases be made in such a way that no information traceable to a specific individual[1] will be disclosed.

The major questions addressed during the year and reported here are as follows:

—What is the nature of statistical disclosure?

—How pervasive a problem is it?

—How can agency requirements be translated into specific disclosure-avoidance techniques?

—How can agency requirements be met without unduly restricting data releases?

—How do agency disclosure-avoidance practices affect data subjects and data users?

1. The Nature of Statistical Disclosure

The problem of statistical disclosure is certainly not a new one. It has long been recognized that any available tabulation of the characteristics of a population is likely to narrow the range of uncertainty about the characteristics of specific individuals known to be members of that population. Recognition of the problem has been heightened by the widespread use of computers and microdata files as well as the increased demand for more detail in statistical releases. The sheer number of characteristics available about a given statistical unit in microdata form, which sometimes produces unique configurations, may make identification possible, even though identifiers (such as names, social security numbers, or employer numbers) have been removed.

Nevertheless, we discovered that comparatively little is known about disclosure. To begin with, there is no widely accepted definition or typology of "disclosure." Probing the definitional issue, we reviewed prevailing statutes, regulations, and policy directives at the Federal level to see what light they might shed on the nature of disclosure. Published literature on the topic was also consulted. Tore Dalenius, consultant to the Statistical Policy Division, OMB, developed a formal definition while working with the Subcommittee. We adopted this definition, as it was judged to provide the best basis for a comprehensive discussion of the disclosure issue. The definition is presented in Chapter II along with the above-mentioned reviews. Citations to the literature appear in Appendix D.

2. Pinpointing Disclosure Potentials and Disclosure-Avoidance Techniques

The definitional effort was augmented by an examination of different types of disclosure and a review of the various factors affecting the potential for unintentional disclosure. Since the nature of the disclosure problem varies significantly for tabulations and microdata tapes, the discussion proceeds separately for the two modes of data dissemination in Chapters III and IV respectively. The latter portion of each of these chapters identifies and describes disclosure-avoidance techniques appropriate for the respective mode of release. To augment this general description, we assembled a description of the disclosure-avoidance practices of several Federal statistical agencies. These appear in Appendix A.

3. Balancing Confidentiality Requirements Against Societal Needs for Information

We have used the term "disclosure avoidance" to describe efforts to *reduce* the risk of disclosure. The release of any data usually entails at least some element of risk. A decision to *eliminate* all risk of disclosure would curtail statistical releases dras-

[1] Except where otherwise specified, the word "individual" as used in this report is meant to cover all types of reporting units—natural persons, corporations, partnerships, fiduciaries, etc.

tically, if not completely. Thus, for any proposed release of tabulations or microdata, the *acceptability* of the level of risk of disclosure must be evaluated. The use of the term "disclosure avoidance" should not be allowed to obscure the vital significance of such evaluations, or to lead to policies which attempt to *eliminate* disclosure risk completely.

In summary, protection of the confidentiality of information about individuals must be balanced against the legitimate needs of society for information. This "Question of Balance" is discussed in Chapter V.

4. Other Considerations

For the most part, our study was confined to matters internal to Federal agencies. However, at one point in Chapter V this limitation is relaxed to examine the impact of agency disclosure practices upon data subjects and data users.

This report does *not* deal with the issue of releasing data with identifiers, whether such release is intentional or unintentional. Our treatment of disclosure differs from that commonly associated with the Privacy Act of 1974, for example, which treats disclosure as transferring information coupled with identifiers. The conception of disclosure advanced here excludes from consideration many identifier-linked confidentiality issues, such as whether statistical data should be immune from mandatory release for administrative, legislative and judicial purposes. By the same token, the report deals only tangentially with the issue of computer security, ignoring the much-discussed potential for penetration and misuse. A substantial literature on that problem already exists, which this report highlights in Appendix B. The more relevant computer aspect is the possibility of mechanizing the search for disclosure risks and the implementation of disclosure-avoidance techniques. Appendix C reports on the development of an automated system to avoid disclosure in tabulations published by the Bureau of the Census from its economic censuses.

5. Findings and Recommendations

Our findings and recommendations appear in Chapter VI. In framing recommendations, we have been mindful of the diversity of statistical activity within the Federal establishment, as well as the complexity of the matter, and refrained from advocating overly generalized solutions. Yet, because we were also mindful of the pressing nature of the disclosure problem, the report includes a number of suggestions for the development and review of agency disclosure-avoidance practices.

B. Auspices

The report represents the collective efforts of the Subcommittee on Disclosure-Avoidance Techniques of the Federal Committee on Statistical Methodology which operated under the auspices of the Office of Federal Statistical Policy and Standards, Department of Commerce (previously the Statistical Policy Division, Office of Management and Budget). The group was originally formed in early 1976 as one of two working groups of a Subcommittee on Confidentiality Issues chaired by Thomas B. Jabine. The working groups were subsequently given separate subcommittee status. The other group, the Subcommittee on Matching Techniques, examined methodological issues associated with the merger of microdata from different data sets.

The opinions expressed here reflect the collective judgment of the Subcommittee and do not necessarily reflect those of the Federal Committee on Statistical Methodology or the Office of Federal Statistical Policy and Standards.

C. Dissemination of Report

This report is intended for circulation among managerial and technical staff of statistical agencies and those Federal offices which release information for statistical and research purposes. The report is intended to apprise such staff more fully of the disclosure problem and encourage appropriate disclosure-avoidance practices at the individual agency level. In addition, we hope this report will furnish the basis for an informed discussion of the disclosure problem within the Federal establishment generally as well as between the Federal Government and its data suppliers and users. It may also be of more general use to persons interested in issues related to the avoidance of statistical disclosure.

CHAPTER II

Defining Statistical Disclosure

A. References in Statutes, Regulations, and Policy Statements

The first requirement of Federal agency policies for avoiding disclosure in the release of tabulations and microdata is that these policies conform with relevant statutes and regulations. In addition, there have been several recommendations on this subject by advisory groups, which, while not binding, often carry considerable weight. This section of the chapter presents and reviews relevant sections of statutes, regulations and reports of advisory groups.

1. The Privacy Act of 1974

The Privacy Act (P.L. 93–579, 1974) does not address the issue of disclosure in tabulations; however, it does have one provision relating to disclosure of microdata. Section 552a(b)(5) provides for disclosure without consent of the individual to whom the record pertains "to a recipient who has provided the agency with advance adequate written assurance that the record 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."

The OMB Guidelines for Privacy Act Implementation (U.S. Office of Management and Budget, 1975) explain the statutory language as follows: "The use of the phrase 'in a form that is not individually identifiable' 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.)" The Guidelines go on to say "Fundamentally, agencies disclosing records under this provision are required to assure that information disclosed for use as a statistical research or reporting record cannot reasonably be used in any way to make determinations about individuals."

Unfortunately, the applicability of this provision of the Privacy Act to the release of microdata from Privacy Act record systems is far from clear. It can be argued that records meeting the requirements of 552a(b)(5), are in general *required* to be released in response to Freedom of Information (FOI) Act (P.L. 93–502, 1974) requests, since they do not come under any of the FOI exemptions. Surely, since all reasonable possibility of identification by recipients is presumed to have been eliminated, such records would not come under 552(b)(6) of the Freedom of Information Act, which exempts from mandatory FOI disclosure "personnel and medical files and similar files the disclosure of which would constitute a clearly unwarranted invasion of personal privacy."

The Privacy Act itself provides in Section 552a (b)(2) for disclosure without consent where such disclosure would be "required under Section 552 of this title" (section 552 is the Freedom of Information Act), and it would seem that most disclosures of information meeting the requirements of 552a(b)(5) of not being individually identifiable would fall under 552a(b)(2) and not 552a(b)(5).

If the above analysis is found to be confusing, this is indicative of the dilemma facing the Federal agency official trying to determine whether and under what conditions the Privacy Act permits him to release a specified microdata file.

2. The Freedom of Information Act

In thinking about disclosure-avoidance policies, it is important to keep in mind that FOI requires Federal agencies to make any records or documents in their possession available to individuals on request, unless such materials come under one of the 9 exemptions in the act. Thus, FOI requests for existing statistical tabulations and microdata files can be denied only if one or more of these exemptions applies. Furthermore, denials in such cases are not *required* by FOI; the materials may be released unless prohibited by another statute or regulation. Three of the 9 exemptions are pertinent, and are discussed below.

Exemption (3).—This exemption formerly referred to matters "specifically exempted from disclosure by statute." However, the Government in the Sunshine Act (P.L. 94–409, 1976) has changed this exemption (effective March 14, 1977) to read "specifically exempted from disclosure by statute (other than Section 552(b)[1] of this title), provided that such statute (A) requires that the matters be withheld from the public in such a manner as to leave no discretion on the issue, or (B) establishes particular criteria for withholding or refers to particular types of matters to be withheld." The effect of the change was to substantially narrow the applicability of this exemption. Agencies, including for example the Social Security Administration, whose confidentiality statutes do not meet the new requirements of exemption (3) now have to rely on one of the other FOI exemptions when they wish to protect statistical tabulations or microdata files from mandatory release under FOI.

Exemption (4).—This exemption refers to "trade secrets and commercial or financial information obtained from a person and privileged or confidential." The extent of applicability of this exemption to statistical tabulations and microdata is not well defined at this time, and will only become clearer as court decisions rule on its applicability to FOI requests for such data.

Exemption (6).—This exemption refers to "personnel and medical files and similar files the disclosure of which would constitute a clearly unwarranted invasion of personal privacy." As in the case of exemption (4), the extent of applicability of this exemption to tabulations and microdata is not yet clear. Recent court decisions have tended to limit its applicability.

3. Agency Statutes and Regulations

Following is a review of selected provisions of agency statutes and regulations relevant to the release of statistical tabulations and microdata. It is *not* intended that this be a full review of agency confidentiality statutes and regulations. We cite here only those provisions which appear to be directly relevant to the question of defining statistical disclosure.

a. *Bureau of the Census, Title 13.*—The relevant portion prohibits the Census Bureau from making "any publication whereby the data furnished by a particular establishment or individual under this title can be identified."

b. *Internal Revenue Service.*—The section of the Internal Revenue Code dealing with "Statistical Publications and Studies" as amended by the Tax Reform Act (P.L. 94–455, 1976) provides that "Nc publication or other disclosure of statistics or othei information required or authorized by subsection (a or special statistical study authorized by subsectior (b) shall in any manner permit the statistics, study or any information so published, furnished, or otherwise disclosed to be associated with, or otherwise identify, directly or indirectly, a particular taxpayer." [2,3]

c. *Social Security Administration.*—Regulatior Number 1, promulgated under Section 1106 of the Social Security Act, deals with "Disclosure of Officia Records and Information." Until recently, Sectior 401.3(k) of Regulation 1 provided that "Statistica data or other similar information not relating tc any particular person which may be compiled fron records regularly maintained by the Department ma be disclosed when efficient administration permits."

d. *Law Enforcement Assistance Administration* —The Crime Control Act of 1973, in Section 524(a) provides that "Except as provided by Federal lav other than this title, no officer of the Federal Government, nor any recipient of assistance under th provisions of this title shall use or reveal any researcl or statistical information furnished under this titl by any person and identifiable to any specific private person for any purpose other than the purpose fo which it was obtained in accordance with this title."

The regulations implementing this Act (Law Enforcement Assistance Administration, 1976) defin "information identifiable to a private person" a "information which either—.

(1) Is labelled by name or other personal identifiers, or

(2) Can, by virtue of sample size or other factor: be reasonably interpreted as referring to a particula private person."

e. *National Center for Health Statistics.*—Publi Law 93–353, Section 308(d) provides that "No information obtained in the course of activities undertaken or supported under Section 304, 305, 306, c 307 may be used for any purpose other than the purpose for which it was supplied unless authorize

[1] The section which sets forth the FOI exemptions.

[2] This section became effective January 1, 1977.

[3] Subsection (a) authorizes annual or more frequent publication "statistics . . . with respect to the operations of the internal reven laws." Subsection (b) authorizes the performance of "special stat tical studies and compilations involving return information" f others on a reimbursable basis.

[4] Passage of the Government in the Sunshine Act referred to earl brought about the need for substantial revision of Regulation Pending final adoption of the revised Regulation 1, the Social Secur Administration is operating under an interim version which does n deal explicitly with this question.

under regulations of the Secretary; and (1) in the case of information obtained in the course of health statistical activities under Section 304 or 306, such information may not be published or released in other form if the particular establishment or person supplying the information or described in it is identifiable unless such establishment or person has consented . . . ."

The common element in these and other agency statutes and regulations is the prohibition of the release of information that can be associated with or identified to a particular statistical unit. In some cases the prohibition is limited to information about private individuals; in others, it extends to information for legal persons, such as businesses.

4. Advisory Committee Reports

a. *The President's Commission on Federal Statistics (1971).*—Recommendations on privacy and confidentiality appear in Chapter 7 of the Commission's Report. Recommendation 7-4 says, in part, "use of the term 'confidential' should always mean that: a. Disclosure of data in a manner that would allow public identification of the respondent or would in any way be harmful to him is prohibited."

b. *The HEW Secretary's Advisory Committee on Automated Personal Data Systems.*—Chapter 6 of the Committee's Report (U.S. Department of Health, Education, and Welfare, 1973) deals with "Special Problems of Statistical-Reporting and Research Systems." In this chapter, the Committee recommends new Federal legislation protecting against compulsory disclosure. One of the features recommended for the legislation was: "The protection should be limited to data identifiable with, or traceable to, specific individuals. When data are released in statistical form, reasonable precautions to protect against 'statistical disclosure' should be considered to fulfill the obligation not to disclose data that can be traced to specific individuals."

A footnote to this paragraph provides a definition of statistical disclosure from an article by Fellegi (1972). "This is a risk that arises when a population is so narrowly defined that tabulations are apt to produce cells small enough to permit the identification of individual data subjects, or when a person using a statistical file has access to information which, if added to data in the statistical file, makes it possible to identify individual data subjects."

c. *The American Statistical Association Ad Hoc Committee on Privacy and Confidentiality (1977).*—The Committee's report includes several recommendations on "Release of statistical summaries and microdata without identifiers." The first of these recommendations is:

"1. General public releases of statistical summaries and microdata files based on either administrative or statistical data sources should be permitted without restrictions or conditions provided that:

(a) All identifying particulars, such as name, address and Social Security number, have been removed, *and*

(b) It is virtually certain that no recipients can identify specific individuals in the files."

For microdata files which do not meet condition (b) of this recommendation, the Committee recommends release for research and statistical purposes only under certain conditions, one of which is that the recipient agrees "Not to release any tabulations or other information that would make it possible for others to identify specific individuals."

d. *The Privacy Protection Study Commission (PPSC).*—The Commission's final report was issued in July 1977 (PPSC, 1977). Chapter 15, entitled "The Relationship Between Citizen and Government: The Citizen As Participant in Research and Statistical Studies," includes several recommendations and policy guidelines relating to the collection, use and disclosure of information about individuals (natural persons) in "individually identifiable form" for research and statistical purposes.

The report defines "individually identifiable form" as "any material that could reasonably be uniquely associated with the identity of the individual to whom it pertains" (PPSC, 1977:572). Thus, it is clear that the Commission was fully aware of the problem of statistical disclosure, and, in fact, in a section of Chapter 15 on "Procedures to Protect Confidentiality" (PPSC, 1977:583–7), there are brief references to the work of this Subcommittee and to several of the disclosure-avoidance techniques discussed in this report.

Recommendation (6) in Chapter 15 (PPSC, 1977: 587) is "That the National Academy of Sciences, in conjunction with the relevant Federal agencies and scientific and professional organizations, be asked to develop and promote the use of statistical and procedural techniques to protect the anonymity of an individual who is the subject of any information or record collected or maintained for a research or statistical purpose."

The text immediately preceding this recommendation makes it clear that techniques to avoid statistical disclosure (at least in its "exact" sense) are intended to be included in the recommended program of activities by the Academy and other organizations.

B. Evaluation of Statutory Requirements

Statutory prohibitions on disclosure are expressed in absolute terms. Thus, the Privacy Act refers to disclosure of a record "in a form that is not individually identifiable." The Census Title 13 prohibits "any publication whereby the data furnished by a particular establishment or individual under this title can be identified."

If these statutory restrictions were interpreted literally, the flow of statistical data from the Federal Government would be stopped or drastically reduced. In a broad sense, *any* release of statistical tabulations reveals some information, at least in an approximate or probabilistic sense, about every individual known to be included in those tabulations. When a microdata file containing numerous items of information about each individual is released, it is virtually certain that many of the records will display combinations of characteristics not possessed by more than one individual in the population, and therefore will be *potentially* identifiable through matching with data that might be available from other sources.

In practice, what is clearly expected on the part of agencies releasing statistical data is an effort to keep the *probability* of disclosure, however defined, at a very low level. Three of the advisory groups cited above confirm this view of the question. Thus, the HEW Committee called for "reasonable precautions to protect against 'statistical disclosure' "; the ASA Committee recommended unrestricted release when "it is virtually certain that no recipients can identify specific individuals in the file."; and the Privacy Protection Study Commission used the word "reasonably" in defining "individually identifiable form." We may also note that the LEAA regulation uses the word "reasonably" in this context whereas the statute did not include any such qualifying term.

This interpretation of statutes, regulations and recommended policies which prohibit disclosure leads to an important conclusion, i.e., *that they do not in themselves provide a clear basis for deciding in any particular case whether data should or should not be released.* The decision on release calls for more specific rules and guidelines. If such rules and guidelines do not exist, then each case will be a judgment call by the responsible official.

A major objective of this Subcommittee has been to determine what rules, guidelines and other criteria are being used by Federal agencies to avoid statistical disclosure; to review and evaluate these materials; and to make its findings widely available for the benefit of statisticians and others who must make decisions on what data to release, and on what terms.

C. Prior Definitions of Statistical Disclosure

We have seen that, without exception, laws and regulations do not provide a sufficiently precise definition of disclosure for operational use in determining what tabulations and microdata files are releasable. We have also reviewed the literature on the subject of statistical disclosure found in journals, reports and other publications. There we have found several attempts at a more precise definition. These are all helpful, but none of them seems to be broad enough to cover all the kinds of statistical disclosure problems met with in practice.

Fellegi (1972) defines "inadvertent direct disclosure (i.d.d.)" as "disclosure of information on an individual who can be identified through his characteristics." He goes on to say that such disclosure "occurs when a user can identify a respondent by recognizing him through his characteristics and learning something about him." In other words, this kind of disclosure only occurs when two things happen:

1. The user *recognizes* an individual member of a population included in a tabulation or microdata file.

2. The user *learns* something about that individual that he did not know from another source.

Many more casual definitions of disclosure include only the first element.

Fellegi does not say whether the information learned must be the exact value of some characteristic, or whether the disclosure can be in the form of a range, or a probability statement about the value in question. Hansen (1971) distinguishes between "exact" and "approximate" disclosure, the latter term being used for the case where a value for a particular individual is disclosed to be within some specified range.

Fortunately, there is now available, in a report by Dalenius (1977) a mathematical treatment of the concept of statistical disclosure which we believe provides an adequate framework for discussion of all aspects of statistical disclosure. Dalenius has kindly agreed to the inclusion of this material in our report.

D. A Proposed New Definition of Statistical Disclosure

The reader is asked to keep in mind that the concept of disclosure presented here is a very broad one. It would not be desirable to require that there be a zero risk of disclosure, as defined below, in any release of tabulations or microdata files. Such a requirement would end a large proportion of all releases now being made. This would be too great a price to pay for complete elimination of any risk of disclosure.

The material which follows in sections D1, D2 and D3 is presented verbatim from Dalenius' report, except for a few changes in terminology to conform with the language and structure of this report.

1. The Insufficiency of Prevailing Definitions

Statistical disclosure is used in the literature in a way which parallels its use in nonstatistical contexts, Thus, in Webster's *Third New International Dictionary*, "disclosure" is defined as:

(1) the act or an instance of opening up to view, knowledge or comprehension.

(2) something that is disclosed.

This definition is, indeed, general; it is by and large consistent with definitions of disclosure in the context of releases of statistical results. An example, Title 13, U.S. Code, Section 9-a-2, gives an implicit definition of disclosure; it states that there shall not be:

> ". . . any publication whereby the data furnished by a particular establishment or individual under this title can be identified."

The definition just quoted is less general than the definition taken from Webster's dictionary, by making *identification* of the object(s) concerned an element of the definition. While this is indeed a crucial difference, it does not make the resulting definition sufficiently specific to serve as a basis for regulations and/or procedures aiming at disclosure control; it does not easily and unambiguously lend itself to implementation.

In sections D2 and D3 an effort will be made to deal with the conceptual problem thus present.

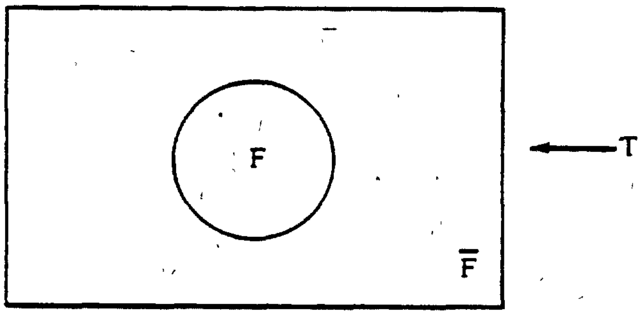
2. A Framework for Defining "Statistical Disclosure"

"Statistical disclosure" is used here in accord with the use of this term in the context of releasing statistics from a survey [5]. In line with this notion of disclosure, the following four components are used to provide the conceptual framework called for:

a. A frame comprising certain objects
b. Data associated with these objects
c. Statistics released from a survey
d. Extra-objective data

(a) The frame

Consider a set of identifiable objects, to be referred to as the total population and denoted by T. In a typical case, T may be "all Swedish citizens." The survey concerns a subset of this total population, *viz.* that subset which is accessible by means of a certain frame; for convenience, this subset will be denoted by F. In a specific case, F may be "Swedish citizens living in Sweden." The complementary subset—i.e., the subset made up by objects in T which are not in F—is denoted by $\bar{F}$. Thus, T is the "union" of F and $\bar{F}$.



In the case of a *sample* survey, it may prove useful to make an additional distinction, *viz.* between objects selected for the sample F_s and those not selected $F_{\bar{s}}$.

(b) Data associated with the objects in the frame

With each object in F, we associate data, which serves three different *functions*:

i. Identifying function:

We will denote the data serving this function by the identifier *I*. In a specific case, *I* may appear as a (registration) number, or as name and street address.

[5] The Dalenius text uses the word "survey" in its broad sense to include a census or other data collection covering the total population. For purposes of this report, the definition may also be applied to the release of statistics based on administrative or program records.

ii. Classifying function:

For purposes of presenting the "details" of the statistics to be released, the objects in F will be associated with certain classes, defined by reference to some classifier C. In a specific case, C may appear as a "code" identifying a subset of F, for example a subset defined with reference to the sex and age of the objects in F.

iii. Information function:

The survey is carried out in order to provide information in terms of certain "survey characteristics" X, Y, . . ., Z. For the object O_J ($J=1, \ldots, N$), the values of these characteristics are denoted by $X_J, \ldots, Z_J$. Typically but not exclusively, these values may be in the nature of *counts* or *magnitudes.*

It may be worth noting that some data may serve more than one of these 3 functions in one and the same survey.

(c) The statistics released from the survey

The *objective* of a survey is expressed in terms of some population and some data C and X, Y, . . ., Z. In order to achieve this objective, the statistics S are released. We will focus on two dierent kinds of statistics:

i. statistics for sets of objects—*"macrostatistics";* typically, the format of a report is used as a means of releasing the statistics.

ii. statistics for individual objects—*"microstatistics";* typically, the format of microdata tape is used as the means of releasing the statistics.

We will elaborate upon the above distinction in sections (1) and (2) below.

(1) Macrostatistics

In the case of macrostatistics, the statistics —counts, magnitudes, etc., as the case may be—concern aggregates of the individual values of the survey characteristics belonging to the respective sets. The following tables are two cases in kind:

Number of beneficiaries by county and age

	Age class				
County	Under 65	65–69	70–74	75 & Over	Total
A ———	3	15	11	8	37
B ———	7	60	34	20	121
C ———	—	4	—	—	4

Average benefit amount by county and age

	Age class			
County	Under 65	65–69	70–74	75 & Over
A ———	$63.30	$94.30	$85.20	$79.60
B ———	62.40	89.90	81.80	72.40
C ———	59.80	92.40	80.40	77.60

These tables—while featuring the characteristics of real life statistics—are admittedly "small."

(2) Microstatistics

In this kind of statistics, the individual values observed with respect to the characteristics X, Y, . . ., Z (possibly in conjunction with the associated classifiers) are released. The identifiers, however, are *not* released. The following excerpt from U.S. Bureau of the Census (1976) is illustrative:

	State of Residence	Urban/Rural	Persons in household	Telephone	Plumbing	Rent	Automobiles	Household type	
Household #1	Virginia*	Urban	3	Yes	Yes	$125	2	h-w family	

	Relationship	Sex	Age	Race	Place of Birth	Years of School	Occupation	Earnings	
Person a	Husband	M	37	W	Kansas	12	Plumber	$13,000	
Person b	Wife	F	35	W	Virginia	12			
Person c	Child	M	6	W	Virginia	1			

Household #2	Virginia	Rural	1	Yes	No	$30	0	Primary Indiv.	

Person a	Primary Indiv.	F	68	N	Alabama	6	Service	$1,400	

Household #3	Virginia	Urban	6	Yes	Yes	$205	2	h-w family	

(d) Extra-objective data

In section (c), we related the *objective* of a survey to two kinds of data: C, and X, . . ., Z, respectively. It is characteristic of the design of a survey that it provides a source of these data.

We will use the term "extra-objective data" to denote any kind of *additional* data; for convenience, these data will be denoted by E. It is characteristic of E that it is not part of the objective of the survey; thus, the design does not explicitly provide a source of these data.

(e) Summary

Thus, the four components of the framework may now be stated as:

i. The frame: F
ii. The data associated with the objects in the frame: *I*, C, X, Y, . . ., Z

* Public Use Sample tapes do not actually contain alphabetic information, but represent the characteristics in the form of numeric codes.

iii. The statistics released from the survey: S
iv. The extra-objective data: E

3. Statistical Disclosure Defined

We will now suggest a definition of disclosure within the conceptual framework presented in section 2.

Thus, consider an object O_K in the total population T. This object may be a member of F, or it may be a member of $\overline{F}$. We introduce a characteristic D which may be one of the survey characteristics X, Y, . . ., Z; or it may be some other characteristic. For the object O_K, this characteristic assumes the value D_K. It is helpful to consider two special cases:

i. $D_K = 1$ if O_K has a certain property otherwise $D_K = 0$.
ii. D_K is measured on a ratio scale: it is expressed as a magnitude.

If the release of the statistics S makes it possibl to determine the value D_K more accurately than i possible without access to S, a disclosure has take place; more exactly, a D-disclosure has taken place In a specific case, this D-disclosure may be an X-disclosure, or a Y-disclosure, etc.

The definition just given applies to both releas of macrostatistics and release of microstatistics. Examples of disclosure for the former case may b found in Chapter III and for the latter case in Chapter IV.

CHAPTER III

Disclosure in the Release of Tabulations (Summary Data) for Public Use

A. The Problem of Disclosure in Tabulations: Typology, Identification and Examples

The problem of disclosure in tabulations will now be discussed. A typology will be listed; ways to identify the various types of disclosure, together with appropriate examples, will be provided.

The definitions of different kinds of disclosure used in this section are very broad. Not all of these kinds of disclosure need necessarily be avoided in all tabulations. The issues involved in determining what kinds of disclosure are acceptable in a particular situation are discussed in section B2 of this chapter.

Our study of the literature on this subject did not reveal any generally accepted definitions of various types of disclosure. The proposed classifications which follow represent an effort to develop a comprehensive and logical description of different types of disclosure. Suggestions for improvement will be welcomed.

Disclosure will be studied both for tabulations involving count (frequency) data and for those containing quantity (magnitude) data. Tables 1 and 2 show examples of count data and quantity data, respectively.

Table 1.—*Number of beneficiaries by county and age*

County	Age class				Total
	Under 65	65–69	70–74	75 & over	
A ____	3	15	11	8	37
B ____	7	60	34	20	121
C ____	—	4	—	—	4

Table 2.—*Average benefit amount by county and age*

County	Age class			
	Under 65	65–69	70–74	75 & over
D ____	$63.30	$94.30	$85.20	$79.60
E ____	62.40	89.90	81.80	72.40
F ____	59.80	92.40	80.40	77.60

1. Exact Disclosure

a. *Count data.*—For tabulations involving counts of persons, establishments, etc., exact disclosure is said to occur when a respondent known to be a member of a set (marginal total) can be determined to be a member of a proper subset (cell). For the disclosure to be exact, this proper subset or detail cell must be defined as narrowly as possible. The detail cell must consist of respondents all having one of the basic, elementary values available from the records of the characteristic defining the cell—single year of age, nearest dollar amount of benefit, a single race category, etc. Table 3 shows that all beneficiaries in County B are black—an example of exact disclosure.

Table 3.—*Number of beneficiaries by county and race*

County	Race			Total
	White	Black	Other	
A ____	15	20	5	40
B ____	0	30	0	30

On the other hand, the inference from Table 4 that no beneficiary in County B is white is not called exact disclosure because the subset of black or other beneficiaries is not as narrowly defined as possible from the records on which the tabulation is based.

Table 4.—*Number of beneficiaries by county and race*

County	Race			Total
	White	Black	Other	
A ____	15	20	5	40
B ____	0	28	2	30

Similarly, the fact that the ages of all beneficiaries in County C of Table 1 can be restricted to the interval 65-69 does not constitute exact disclosure as defined here because the age interval defining the detail cell does not represent a single year of age.

In summary, exact disclosure from count data can be identified as follows: A marginal total (in the dimension n–1) of an n-dimensional cross tabulation equals one of its detail cells; this detail cell is as narrowly defined as possible.

b. *Magnitude data.*—Exact disclosure from magnitude data can occur as a result of the publication of the value of a quantity corresponding to a cell with only one member. For example, the total sales for the single establishment in Industry B is disclosed by Table 5.

Table 5.—*Total sales, by industry*

Industry	No. of establishments	Total sales
A ____	18	$450,000,000
B ____	1	125,000,000

A second type of exact disclosure from magnitude data occurs when auxiliary information concerning the possible numerical values of the characteristic under consideration can be used to determine the exact quantity for every member of a given cell. For example, consider the situation presented below:

Table 6.—*Average monthly benefits, by State*

State	No. of beneficiaries	Average Monthly benefit
A ____	4	$158
B ____	36	190

If the maximum possible monthly payment to any beneficiary under the program studied in Table 6 is $190, then the user will know that each person in State B receives precisely $190. However, the exact value of the payment to any beneficiary in State A is not disclosed.

In summary, exact disclosure of the first type from quantity data is identified by the publication of the numerical value of a characteristic corresponding to a cell with one member. Exact disclosure of the second type from magnitude data is identified by the following equalities:

$$A = L, \text{ equivalently } T = LN$$

or

$$A = U, \text{ equivalently } T = UN,$$

where

A is the average and T is the total value among all N members in a cell, U and L are the maximum and minimum possible values, respectively, for any member in the cell.

2. Approximate Disclosure

a. *Count data.*—When all members of a total belong to one detail cell, the disclosure is approximate if the detail cell is not as narrowly defined as possible: otherwise, the disclosure is exact.

When all members of a total can be restricted to a proper subset of detail cells, there is approximate disclosure because it is disclosed that no member of the marginal total belongs to any of the empty cells.

Table 1 allows the user to restrict the age of each beneficiary in County C to the interval [65, 69]. Table 4 does not exactly specify the race of any person, but it shows that the race of each beneficiary in County B is either black or other, not white.

Both of the above examples illustrate approximate disclosure from count data.

Approximate disclosure from count data can be defined and identified as follows: A marginal total (in the dimension n–1) of an n-dimensional cross tabulation equals one of its detail cells, or the sum of a proper subset of detail cells (equivalently, the value of one or more detail cells is zero); but the disclosure is not exact.

b. *Magnitude data.*—In a broad sense the publication of a figure for quantity always permits the user to estimate, however crudely, the value of a characteristic corresponding to a given member of the cell. For example, the monthly benefit for each of the four beneficiaries in State A of Table 6 must be less than $632. Further, the total sales of each establishment in Industry B of Table 7 can be placed inside the interval [0, 125,000,000].

Table 7.—*Total sales, by industry*

Industry	No. of establishments	Total sales
A ____	18	450,000,00
B ____	5	125,000.00

Often, the information provided in cases such a the above will not be sufficiently accurate or sensi tive to require corrective measures. However, if th number of members in the cell is sufficiently smal the interval of possible values for the quantity asso ciated with a particular individual will be narrow enough to be considered a disclosure problem (Co 1976).

With the assumption that all values for quantit are non-negative, the interval of possible values c a characteristic for a particular cell member is [0 T] if the total, T, is published; equivalently, th interval is [0, NA] if the average, A, and cell siz N are published.

Sometimes auxiliary information obtained fro sources external to the summary data under consic eration can enable the user to estimate the value of an unpublished quantity more accurately. For example, if an employment distribution shows that all establishments in Industry B of Table 7 have approximately the same number of employees, the user can estimate a value $25,000,000 for the sales of each establishment. In the same vein, if it is known from another data source that the largest establishment of the five employs 80 percent of all workers in Industry B, a reasonable estimate for total sales for that establishment would be $100,000,000.

In some situations, auxiliary information admitting more accurate approximation to values of aggregate data can be obtained from external sources other than statistical tabulations. In particular, legal requirements used in conjunction with summary data may determine narrow upper and lower limits for the value of a quantity for an individual respondent.

For example, in Table 6 if the maximum benefit is $192, then it can be shown that each individual person in State B must receive at least $120—a restriction of each beneficiary's payment inside a range of values unknown prior to publication of the data.

In general, if maximum and minimum values of the characteristic in question are known, such disclosure will occur under the following conditions:

$$\text{Either } A < L + P\left(\frac{U-L}{N}\right)\text{, equivalently}$$

$$T < LN + P\,(U-L)$$

$$\text{or } A > U - P\left(\frac{U-L}{N}\right)\text{, equivalently}$$

$$T > UN - P\,(U-L) \text{ hold,}$$

where A is the average and T is the total value among all N members in a cell, where $N > 1$; U and L are the maximum and minimum possible values, respectively, for any member in the cell; and P, where $O < P < 1$, specifies the relative size of the interval chosen to define disclosure of the value of the characteristic under consideration. For example, if disclosure is defined as knowing that the value for an individual lies within a quarter of the range (U-L) then $P = .25$.

Finally, in some instances better approximations for the quantity data of an individual respondent can be computed by a user with precise information about a subset of members of the cell. This type of disclosure is discussed later in this chapter (see A 5: "Internal Disclosure") and in Appendix C.

3. Probability-Based Disclosures (Approximate or Exact)

Sometimes although a fact is not disclosed with certainty, the published data can be used to make a statement which, within the framework of an implied probability model, has a high probability of being correct. For example, in Table 8 it is very likely that a given beneficiary in County B has a monthly income in excess of $2,000.

Table 8.—*Monthly income of beneficiaries*

County	Number of persons with income		
	Under $1000	$1000–$2000	Over $2000
A ____	70	60	65
B ____	10	20	230
C ____	30	50	40

Similarly, from Table 4, in the absence of other information, we might assign a probability of 0.93 that a person known to be a beneficiary in County B is black.

Identification of probabilistic disclosure can be described as follows:

$$D < SP_1 \text{ or } D > SP_2$$

where

D is the number of members in the detail cell,

S is the number of members in the total cell,

P_1 is the smallest permissible proportion of members in a detail cell among all members belonging to the marginal total, and

P_2 is the largest permissible proportion of members in a detail cell among all members belonging to the marginal total.

As was the case for approximate disclosure for aggregates, the appropriate values of P_1 and P_2 in a particular case must be determined by the agency releasing the tabulations. In many cases, the agency may not consider it necessary to avoid probabilistic disclosure at all; in such cases, we would set $P_1 = 0$ and $P_2 = 1$.

4. Indirect Disclosure

Up to this point, the examples concerning exact, approximate, and probabilistic disclosure have involved information provided directly by published figures. This type of disclosure is said to be direct.

However, information can often be derived by algebraic manipulation and/or logical operations performed upon data obtained from different tables based on the same data. If the publication of a derived figure would result in one of the types of disclosure discussed above, then indirect (exact, approximate, or probabilistic—whichever is appropriate) disclosure is said to occur.

Table 9.—*Number of persons with hospital and medical coverage, by age and sex*

Age	Hospital & Medical coverage		
	Male	Female	Total
Under 65	1,714	1,820	3,534
65–74	1,517	1,630	3,147
75 and over	1,402	1,510	2,912
Total	4,633	4,960	9,593

Table 10.—*Number of persons with medical coverage, by age and sex*

Age	Medical Coverage		
	Male	Female	Total
Under 65	1,719	1,829	3,548
65–74	1,519	1,630	3,149
75 and over	1,402	1,510	2,912
Total	4,640	4,969	9,609

Neither Table 9 or Table 10 discloses individual information directly. However, by application of algebraic and logical operations to both tables, it follows that all men 75 and over with medical coverage have hospital coverage; all women with medical coverage but without hospital coverage are under 65, etc.

As a further illustration of indirect disclosure, suppose Industry A consists of two disjoint subindustries A1 and A2, and that the following information is available from various tables.

Industry	No. of Companies	Total sales
A	5	$200,000,000
A1	4	150,000,000

By subtraction, the total sales of $50,000,000 is computed for the one company belonging to Industry A2.

To identify indirect disclosures, a determination must be made to see if a logically defined but unpublished cell, which would itself constitute a disclosure, can be derived from published cells. Because data from all sources available to the user must be considered, this work can get quite involved. Discussions of this complex problem are given by Cox (1976) and Fellegi (1972).

5. External or Internal Disclosure

Almost all of the above discussion has centered upon external disclosure, i.e., disclosure to someone who is not a member of the tabulated cell. Attention will now be focused upon internal disclosure—that is, the situation in which members of a group use their own as well as published data to obtain confidential information about others in the group. When some members of a group collaborate for this purpose, we will refer to this subset as a "coalition."

Table 11 furnishes an example of internal disclosure for count data. The black worker in County C can determine from the table that every other employee in his industry and county is white.

Table 11.—*Race of workers in industry A, by county*

County	Total	White	Black
A	144	132	12
B	238	138	100
C	94	93	1

If there were precisely two black workers in County C instead of one and if they knew each other, they could deduce that all other employees in their industry and county are white.

If the maximum possible benefit for each of the beneficiaries of Table 12 were $140, it would be impossible for a user not belonging to County B to determine the payment to either person in that county. However, either beneficiary could readily compute the payment to the other person by use of the published cell.

Further, if one person in County A of Table 12 received a benefit of $40, he would know that each of the other persons must receive between $120 and $140.

Table 12.—*Number of beneficiaries and average payment amount*

County	Number	Average Payment Amount
A	3	$100
B	2	70

Another example of internal disclosure from quantity data is given by Table 7 which was also discussed in conjunction with approximate disclosure. As previously mentioned, by subtracting the value of its own sales from the published value $125,000,000 an establishment can estimate the value of sales for its competitors with greater accuracy, perhaps, than they would like.

Finally, internal probabilistic disclosure can be discussed by modifying data for County C of Table 11 as follows:

Total	White	Black
94	92	2

If either black employee knows that Mr. X is in his industry and county, the probability is only 1/93 that Mr. X is black.

For the sake of completeness and summarization, the following list is provided for the identification of the different types of internal disclosure. Definitions are analogous to the corresponding ones for external disclosure.

a. *Count data (direct or indirect disclosure).*—The potential for internal disclosure is affected by two new factors not relevant to external disclosure. The first is the maximum size of coalition against which protection is believed to be necessary; the second is the distribution of the coalition members among the data cells to be protected. Since there is usually no way of knowing what the distribution of any particular coalition might be, the conservative approach in all cases is to protect against the distribution that would result in the greatest degree of disclosure.

In the discussion below,

S is the published number of members in the total cell,

D is the published number of members in a detail cell,

C is the maximum coalition size for which protection from disclosure is considered necessary, and

X is the number of coalition members also belonging to the detail cell.

Note that the number, X, of members of a coalition of size C which belong to a detail cell of size D must satisfy the following:

$$0 \leq X \leq \text{minimum } (C, D).$$

(1) *Exact disclosure:* The difference between the values of a marginal total and one of its detail cells is equal to the number of members of a coalition not belonging to the detail cell (equivalently, $S-D = C-X$), the detail cell is as narrowly defined as possible. In a plan to guard against such disclosure by coalitions of size C, the extreme case $X = 0$ must be considered; that is, $S-D \leq C$ should be avoided in publications.

(2) *Approximate disclosure:* There exists at least one non-empty detail cell entirely contained in a coalition, but the disclosure is not exact. For this detail cell we have $X = D$. In a plan to guard against such disclosure by coalitions of size C, $D \leq C$ should be avoided in publications.

(3) *Probabilistic disclosure:*

(i) $D-X < (S-C)\ P_1$,

where D, X, S, and C are as defined previously and P_1 is as defined for external probabilistic disclosure. In a plan to guard against such disclosure by coalitions of size C, the extreme case $X = C$ must be considered; that is, $D-C < (S-C)\ P_1$ should be avoided in publications.

(ii) $D-X > (S-C)\ P_2$,

where D, X, S, and C are as defined previously and P_2 is as defined for external probabilistic disclosure. In a plan to guard against such disclosure by coalition, of size C, the extreme case $X = 0$ must be considered; that is, $D > (S-C)\ P_2$ should be avoided in publications.

b. *Magnitude data (direct or indirect disclosure).*—

(1) Exact disclosure: After a coalition of size C adjusts a published figure by means of its own data, the revised value involves either type of exact disclosure for magnitude data described for the external use. Equivalently, a quantity is published for a cell of size C+R, containing a coalition of size C, where one of the following conditions holds:

(i) $R = 1$

(ii) The revised value of the published figure, obtained by adjusting for the contribution of the coalition, is a maximum or a minimum possible value determined from external, auxiliary information as described on page 12.

(2) *Approximate disclosure:* With an adjustment of a published quantity figure by use of information about itself, a coalition of members of a cell can estimate, more accurately than an outside user, a quantity value corresponding to a member of the cell outside the coalition.

For example, two beneficiaries, each receiving a monthly benefit of $250 in State A of Table 6 would know that each of the other two beneficiaries must receive less than $132.

Given that the (unpublished) values for sales in Industry B of Table 7 are as shown below:

Establishment	Sales
1	1,000,000
2	1,000,000
3	1,000,000
4	22,000,000
5	100,000,000

it follows that establishments 4 and 5 can obtain sensitive and somewhat accurate information about each other (especially if each is aware of the relative sizes of the other four members of the cell). In particular, establishment 5 can deduce that establishment 4 has at most $25,000,000 in sales.

In general, if all quantities are nonnegative, the interval of possible values for a particular cell member outside a coalition is $[0, T - Q_c]$, or equivalently $[0, NA - Q_c]$ where T is the published total, A is the published average, N is the cell size, and Q_c is the value of the quantity for the coalition.

Finally, if upper and lower limits for the possible value of a quantity corresponding to an individual respondent are known, then internal, approximate disclosure can be identified as follows for aggregate data:

$$A < \frac{Q_c}{N} + \left(1 - \frac{C}{N}\right)L + \left(\frac{U-L}{N}\right)P; \text{ equivalently, } \quad T < Q_c + (N-C)L + (U-L)P$$

or

$$A > \frac{Q_c}{N} + \left(1 - \frac{C}{N}\right)U - \left(\frac{U-L}{N}\right)P; \text{ equivalently. } \quad T > Q_c + (N-C)U - (U-L)P,$$

where

A is the published average and T is the published total value for all N members in the cell,

U and L are the maximum and minimum possible values, respectively, for any member in the cell,

P, $0 < P < 1$, specifies the relative size of the interval which defines disclosure of the value of the characteristics under discussion,

C is the number of members in the coalition, and

Q_c is the unpublished value of the quantity corresponding to members of the coalition.

(3) *Dominance rules and their relation to internal approximate disclosure of magnitudes:* Cell suppression is commonly used as a technique to avoid exact and approximate disclosures in tabulations of magnitude data. Typically, "dominance rules" are established to determine which cells should be suppressed. These rules are of the following general type:

> If n or fewer units account for p percent or more of the cell total, the cell must be suppressed.

For example, we might say that if 1 or 2 firms account for 80 percent or more of total sales in a particular cell, that cell should not be published. One consequence of such a rule would, of course, be to require that all published magnitude cells be based on data for 3 or more firms.

The effect of dominance rules is to limit the precision with which magnitudes for individual units can be estimated from the published data by persons who have exact or approximate knowledge of values for one or more members of the cell. In particular, these rules limit the extent of internal approximate disclosure of magnitude data, as defined earlier in this chapter.

Further discussion of dominance rules and their relation to approximate disclosure appears in Appendix C.

If a dominance rule is used to determine when a cell magnitude should not be published, knowledge of the exact rule can make it possible for a member of the cell to obtain more accurate information about his competitors than would otherwise be the case. This may readily be understood from an example.

Suppose a published cell shows sales for 1976 of $1,000,000 for 6 companies in a particular industry. Company A knows that its own sales in 1976 were $750,000. If Company A does not know the dominance rule, it can deduce only that none of the other 5 companies had sales of more than $250,000. If the dominance rule is published however, additional information may be available to Company A. Consider two possibilities:

1. The rule is that no cell is published if 1 or 2 companies account for more than 90 percent of the total. In this case, Company A will know that none of its competitors had sales of more than $150,000.
2. The rule is that no cell is published if 1 or 2 companies account for more than 80 percent of the total. In this case, Company A will know not only that none of its competitors had sales of more than $50,000, but also that each of the 5 other companies had sales of *exactly* $50,000 (since 5 companies must account for sales of $250,000, and none of them can have sales of more than $50,000).

B. Evaluating the Disclosure Problem

The definition of statistical disclosure adopted for this report is, as mentioned earlier, very broad While it may not be feasible to try to avoid completely the possibility of disclosure, it is imperative to exercise disclosure control. Doing so calls for an evaluation as to (1) the level of risk of disclosure inherent in a proposed publication; (2) the acceptability of that risk; and (3) the assurances given to persons (data subjects or others) who provided the information. In what follows, we will address these three points.

1. The Level of Risk of Disclosure

We will now identify four factors which determine the risk of disclosure. In a real-life situation, it will be necessary to try to evaluate their combined effect.

a. *The relative size of the sample.*—As a first approximation, the risk of disclosure is smaller for tabulations based on a *sample* survey than for tabulations based on a complete survey; and by the same token, the smaller the sampling fraction, the smaller is the risk of disclosure.

This evaluation is reasonable when we are dealing with surveys based on designs characterized by the use of an equal probability of selection method. Many large-scale surveys are of this type. If the overall sampling fraction (usually denoted by n/N) is "small," say less than .05, it is less likely that a disclosure will take place.

If, however, the design does not involve equal probability of selection, the situation is different; in fact, for some types of sampling design, the risk of disclosure may be very great for some large reporting units. As an illustration, consider the total of a characteristic with a highly skewed distribution. An example in kind is a survey to estimate total production. In such cases, an efficient sampling design would call for selecting relatively few small units. Disclosure potential would, therefore, be much higher for the large units than for the small units.

The protection against risk of disclosure afforded by a small sampling fraction is considerably less where particular reporting units are, for whatever reason, known to be members of the sample. For example, if a sample is selected based on ending digits of social security numbers, the risk of disclosure is clearly greater if the digital sampling patterns actually used to select the sample are known.

Similarly in a two-stage sample, if the identities of the primary units in the sample are known, then the sampling fractions within these primary units, rather than the overall sampling fraction, determine the degree of protection against the risk of disclosure. More generally, in multi-stage samples, protection is a function of the sampling fractions *within* units known to be in the sample.

b. *The detail provided in the tabulation.*—A publication which provides only "overall" estimates is less likely to generate large risks of disclosure than a publication which provides detailed breakdowns of these estimates.

It is useful to make a distinction between two kinds of breakdowns, *viz.*, (1) by geography, and (2) by other classifiers.

If the data are presented for very *small* areas, the risk of disclosure is typically larger than for large areas. It is this experience which underlies the rules used by the Census Bureau to provide less detailed tabulations for areas such as census tracts and city blocks than it does for large areas such as SMSA's.

If data are published for small "cells" identified in terms of other classifiers such as age, sex and race (perhaps in combination with geography), the risk of disclosure may be large: the smaller the cell, the larger the risk.

c. *The quality of the data.*—If the data on which estimates are based are impaired by non-sampling errors, the risk of disclosure is smaller than in the case of more accurate data. This is in fact why "noise" is sometimes intentionally introduced into estimates.

d. *Availability of external information.*—The existence of external information—for example, information available through directories or other institutional records—may make the risk of disclosure significantly higher than it would be if that information were not available.

In a real-life situation, the survey statistician should, when planning the survey, take these and other factors into account; to some extent, the risk of disclosure can be controlled by the proper choice of survey design. This type of control must, however, be supplemented by disclosure analysis of the proposed publication.

2. The Acceptability of the Disclosure Risk

The crucial point of the disclosure analysis just referred to is to determine if a certain risk of disclosure is too high or too low. It is too high if it may cause non-negligible harm to an individual being subject to disclosure, or to the statistical agency by impairing its ability to collect data in the future. It is too low if it unnecessarily reduces the amount of useful information that can be provided.

Three factors which may be considered in an effort to determine whether a certain disclosure risk is acceptable or not are listed below.

a. *Sensitivity of data.*—Some types of data are clearly more sensitive than others; it suffices to mention data dealing with financial matters, health, sexual behavior, and drinking habits. On the other hand, some data may, at worst, disclose something that is entirely obvious or completely innocuous, or available in public records.

For many data, the degree of sensitivity may be a decreasing function of their age.

b. *Possible adverse consequences of disclosure.*—This topic is closely related to the sensitivity of data. The more sensitive the data are, the more adverse the consequences of disclosure are likely to be.

Clearly the *kind* of consequences caused by disclosure should be taken into account in the disclosure analysis. If the disclosure of some particular datum may reasonably be expected to create a social, economic or legal *problem*, the risk of disclosure must be kept very small. Thus, disclosing that someone has been treated for venereal disease, drinking problems, etc., may generate such a problem.

3. The Assurances Given to the Respondents

Consideration must be given to what assurances have been given to the data subjects or other persons providing information about uses of the data. Under no circumstances should such assurances be violated.

If the information is definitely non-sensitive *and* no promise of confidentiality was given the data subject, then the concern about possible disclosures would be considerably reduced.

C. Disclosure-Avoidance Techniques

A major goal of statistical agencies is to produce and publish as much useful and usable statistics as possible for the benefit of their clients. The need to avoid the unintentional disclosure of sensitive information concerning individual persons or organizations forms a constraint on this endeavor. The statistical agency, therefore, must find or develop techniques that will effectively avoid disclosure while at the same time permitting maximum useful statistical information to be conveyed. The agency would also seek to accomplish this by a method that is both simple and economical.

Techniques for preventing disclosure through statistical tabulations fall into three general classes: data suppression, rolling up data, and disturbing the data.

1. Data Suppression

a. *Cell suppression.*—A data item which, it is determined, could lead to disclosure may simply be suppressed, i.e., the figure is omitted and replaced by an asterisk or other symbol which indicates that the figure is being omitted to maintain confidentiality for the subjects of the table. However, further care must be taken to assure that the disclosing figure may not then be deduced by subtraction, which requires that another figure in the same row and another in the same column also be suppressed, assuming it is desired that no changes be made in the row and column totals. In addition, at least one figure would need to be suppressed—the one at the intersection of the other row and column of the second and third suppressions—to assure that the other suppressions also cannot be deduced by subtraction. Thus, if the row and column marginal totals are to be left unchanged, it is necessary in a two-way distribution to suppress at least four figures to avoid a disclosure.

It is also possible that data in other tables published from the same body of data may enable one to deduce the suppressed figures. Therefore, it is necessary to review all relevant tables to ensure that they do not contain disclosures and also that through a process of subtraction or other algebraic operations they do not enable disclosures to be made, and all necessary suppressions must be made to avoid the possibility of disclosure. Cox (1976) discusses a linear programming technique for exposing cells which require suppression to avoid disclosure.

So as to provide maximum consistency the suppression of certain data items may be made contingent on the acceptability of a "diagnostic" item. For example, in economic censuses if sales in a particular kind of business must be suppressed, then employment, payroll and certain other figures are automatically suppressed with it. This enhances consistency, avoids incidental disclosures, and reduces costs.

b. *Table suppression.*—Many (though not all) disclosure problems can be avoided inexpensively through the elimination of all tabulations involving fewer than some minimum number of cases. Thus, in the 1971 Census of Population in the United Kingdom, no tabulations were presented for enumeration districts having fewer than 25 persons or fewer than 8 households; for such enumeration districts only the total numbers of persons and households were given (Newman, 1975:6). In the 1970 Census, the U.S. Bureau of the Census suppressed distributions by a particular characteristic for any universe in which there were fewer than 5 cases (Barabba and Kaplan, 1975:9). In guidelines for the Social Security Administration (1977) it is suggested that separate tabulations for counties having fewer than 50 beneficiaries be avoided.

For a general discussion of the use of suppression, see Sweden, National Central Bureau of Statistics (1974:32–34). For a discussion of the use of suppression in the U.S. Bureau of the Census, see Barabba and Kaplan (1975:7–10).

2. "Rolling Up" Data

Problems of confidentiality can frequently be solved by changing the structure of tables in such a way that the disclosure possibility is eliminated. Thus, rows or columns can be combined into larger class intervals or new groupings of characteristics. This may be a simpler solution than the suppression of individual items, but it tends to reduce the descriptive and analytical value of the table.

It may also be expensive in that it might require that a few tables be customized in a large set of tables, the remainder of which are produced mechanically in identical formats. General discussions of the rolling-up process are to be found in Sweden, National Central Bureau of Statistics (1974:31–32) and in Social Security Administration (1977:6–7).

An indirect but common example of rolling-up exists in data bases where the Standard Industrial Classification system is used. That hierchical system has 2-, 3- and 4-digit levels providing successively greater detail. When data are suppressed at the 4-digit level the 3-digit level summary provides the benefits of intermediate rolling-up.

Hansen (1971:51) points out that using broad enough class intervals may even avoid approximate disclosure (in the terminology of this report, unacceptable approximate disclosure), for example, when the upper limit of each interval is at least double the lower limit.

3. Disturbing the Data

This process involves changing the figures of a tabulation in some systematic fashion, with the result that the figures are insufficiently exact to disclose information about individual cases, but are not distorted enough to impair the informative value of the table.

Ordinarily rounding is the simplest example. Figures in a table may, for example, be rounded to the nearest multiple of 5. Where the figures involved are very large, this will have little or no effect on the informative value of the tables. If all cells in a table are rounded by the same rules, totals will not always agree with the sums of the detailed cells. If this is considered undesirable, the most detailed cells can be rounded and then added to obtain totals at various levels. Ordinary rounding was used for most tables involving large areas in the 1971 United Kingdom Census (Newman, 1975:9–10). Values of 0, 1, or 2 were replaced by asterisks; percentages were computed from the rounded tables.

There is a growing body of techniques for avoiding disclosure involving the introduction of random error into the figures to be published. For example, in tables relating to small areas prepared from the 1971 United Kingdom Census, to each figure was added, at random, −1, 0, or +1, in the ratio of 1, 2, 1. Enumeration districts were paired, each having opposite correction factors in comparable figures, so that the totalled figures from a set of districts would be accurate, except if there was an odd number of districts in the set (Newman, 1975:3–8).

One possible approach is to introduce "noise" into the file of microdata, thus avoiding the possibility of disclosure in any tabulations produced from the file. This method may simplify matters for the data producer, but it creates problems for the user (Dalenius, 1974).

"Random rounding", a method which has received considerable attention in recent years, combines elements of both rounding and introducing random disturbances. Each figure is rounded to a multiple of some integer, usually 5, but not necessarily to the nearer one. Whether a figure is rounded up or down is determined at random, with the chance of rounding up or down depending upon the amount of change necessary: (Murphy, date unknown: 68–70; Social Security Administration, 1977:7–9).

Final Digit	Probability of Rounding Up
0 or 5	0
1 or 6	1/5
2 or 7	2/5
3 or 8	3/5
4 or 9	4/5

Nargundkar and Saveland (1972) describe and give theoretical support to the use of this method in the tabulations published from the 1971 Canadian censuses of population and housing. Fellegi (1975) presents a technique for controlling the random rounding to assure that the totals will be correct at some predetermined higher geographical area level.

The Swedish Statistical Bureau proposes another random rounding technique which may be used if it is simply desired to remove ones from a table. The one is rounded randomly down to zero with a probability of 2/3 and up to 3 with a probability of 1/3 (Sweden, National Central Bureau of Statistics, 1974:34–35).

The models discussed above for disturbing data are all additive. Multiplicative models are also feasible. Hansen (1971:55–56) suggests one which involves disturbing the figure by a factor within the range of .5 to 1.5, the factor being chosen at random.

4. Limiting Distribution

Situations may arise in which it is not necessary to take special steps to avoid disclosure from statistical tabulations. Under certain conditions a table may be made available to a particular organization, even though the table could not be published for reasons of maintaining confidentiality. An actual example is in the tables on local area social security data provided by the Office of Research and Statistics, Social Security Administration, to the Bureau of Economic Analysis. As a result, the expense of revising the table is avoided, and the actual distribution is available for full research use. This can be done when the receiving organization guarantees (and has the legal authority) to provide fully adequate protections to the confidentiality of the data while it has custody of them.

For one agency to make potentially identifiable data available to another, conditions such as these may be required:

a. The activity must be in accordance with the laws governing the programs of the respective agencies.

b. There must be a legitimate research purpose to be served by the process.

c. The receiving agency must be strictly and legally accountable to the providing agency for its security program.

d. The receiving agency must demonstrate that it has adequate security provisions.

e. The likelihood that any information potentially harmful to an individual would be derived from the tables would, even so, be extremely low.

f. The receiving agency would not and could not be required to turn the data over to any third party, even under subpoena or a Freedom of Information Act request.

g. The providing agency would have opportunity to review any publication of information from the data to insure that no potential disclosures are published.

h. At the conclusion of the project, and no later than some specified date, the receiving agency would either return or destroy all of the tables involved.

i. Significant sanctions or penalties for improper disclosure would apply.

5. Evaluation of Alternative Techniques

If it is determined that there is a possibility that the publication of a table, or a datum within a table, might result in harm to some individual or organization, but, nevertheless, the table has sufficient value that, at least in some form, it should be published, then a decision must be made as to which technique will be used to avoid the disclosure. A number of examples have been cited; various other techniques are also possible. Four principal questions must be weighed in the making of this decision:

a. *The degree of protection provided.*—All of the described methods reduce considerably the likelihood of a disclosure; some give virtually absolute protection against the possibility of disclosure but are more drastic in terms of loss of information.

b. *Effects on users of the data.—All* of the techniques listed have some effect in reducing the value of the data to the user. There is some loss of information inherent whenever data are suppressed, combined, or disturbed. The Swedish method of removing ones from tables by changing them to 0's or 3's perhaps does the least harm to the data conveyed. At the other extreme, the method of "random rounding" to multiples of 5 has considerable effect, since it can cause any figure to be changed by as much as 4. In general, both of these data disturbing techniques may also yield inconsistent figures for the same data items in independently derived totals. Suppression could make some analyses impossible, particularly where the user wants to combine a number of smaller units to obtain totals and other statistics not provided in the tables. The multiplicative method cited by Hansen could cause any figure to be halved or increased by 50 percent. The Swedish suggestion for substituting a range for a sensitive value can also have severe effects if the range is relatively large. Even the smallest of these changes may affect the value of the published data for descriptive or analytic purposes (Dalenius, 1974:220).

With the increasing use of computers in data analysis, particularly where a large number of areas are to be compared, the uniformity of the data input is another factor affecting users. In this context. rolling-up—so that dimensions of the data matrix vary from unit to unit—creates considerable difficulty. Suppression is also problematic in that suppression at any level can prevent the development of a desired total. In this context the data disturbing techniques may be most satisfactory—in that data are always present and they can be added together without biasing effects on the totals derived. Other statistics such as ratios, e.g., persons per household, can be affected; however, with suitable precautions, these effects can be minimized.

c. *The "identifying" nature of the subject items.*—Some subject characteristics are more likely than others to lead to the ability to associate data with a particular individual. A tabulation of race and sex by income probably has more disclosure potential than a similarly detailed table of major field of study in college by income—assuming that race and sex are more readily observable than major field of study. Area of residence is considered highly identifying in nature, and frequently geographic or size of area characteristics are considered separately from any "subject" characteristics of a respondent in disclosure rules. On the other hand opinions recorded in a survey are normally of minimal utility in identifying a respondent.

The Census Bureau, for instance, has in the past used area of residence and race as the critical variables in determining the publishability of small area population census tabulations. If certain minimum population criteria were met in each area, then other characteristics of that population would be provided. On the other hand, the Census Bureau was willing to make available journey-to-work data from the 1970 census in the form of an origin-destination matrix classified by mode (auto, bus, etc.) without any disclosure control, on the assumption that journey-to-work characteristics are highly changeable (the question was asked relative to "last week") for an individual and therefore non-identifying.

d. *Cost.*—Any procedure used to avoid disclosure in statistical tables will involve some cost to the statistical agency. There will be cost in the use of some operating funds, in the use of personnel time that would otherwise be available for other activities, in the computer programming, debugging, and processing, and in time required for the total process and the resulting delay in publication.

* * *

Agencies cited have studied the problem and have tended to settle on one particular technique to be used for all publications of a particular census, or as standard operating procedure. Once this is done and staff understand it, the procedure becomes routinized and automatic. Computer programs are written to automatically "purify" the tables in the system on a mass-production basis, and costs are minimized. All of the techniques described are capable of computerization, and some software packages are available (Cox, 1976:14–15). But such mass procedures may also result in wholesale losses of valuable information. Study of the effects of such procedures may reveal that in many instances the system's application resulted in particular losses of information that are both unfortunate and unnecessary. As described in Appendix C, the Census Bureau has developed programs which attempt to minimize the number of suppressions in magnitude data.

Each statistical agency must make its own study and its own decision to answer this question: How can we do our job of making available the needed data in our area, while at the same time we make sure that no confidential information about any person or any establishment is accidentally released through the tables we publish?

Selected agency policies and practices to avoid unintentional disclosures are noted in Appendix A.

CHAPTER IV

Disclosure in Microdata

A. Nature of the Problem

1. Definition of Microdata

We use the term microdata to refer to files in which each record provides data about an individual person, household, establishment or other unit. An agency's own files of basic records from a survey or other data collection are thus microdata, and normally they are summarized or aggregated to produce statistics for the reports and publications discussed in Chapter III.

Release of microdata to a data user outside the originating agency can serve legitimate and important public purposes in that the data may be useful for many more tabulations or other analyses than the originating agency is prepared to provide. Certain statistical applications (e.g., simulation models) require input in microdata form.

Obviously, release of records about individuals raises the issue of disclosure. Some files are by law not confidential, for example, those from the Census of Governments from which detailed data for specific governmental units are released. On the other hand, most data bases are covered by statutes (discussed in Chapter II) which prohibit the release of data from which information may be gained about identifiable individuals.

Agencies which release microdata for outside use have construed applicable law and regulations to permit the release of individual information insofar as it is not specific enough to allow identification of the individual. Invariably names and addresses, social security numbers and other positive identifiers are removed. Further, certain other information, such as location, is generally withheld or provided only in broad categories.

Microdata is a particularly popular form of release since it gives the user considerable flexibility in his or her analyses. The capacity of data users to perform such analyses has been and is continuing to increase rapidly with the availability of computer resources. At the same time the statistical agency is frequently impelled to release microdata as a labor-saving device—it reduces somewhat the need for extensive published tabulations, and it cuts down on requests for special tabulations which are sometimes seen as diverting agency resources. Thus the dissemination of data in microdata form is steadily increasing.

2. Federal Agency Examples of Microdata Release

a. *Bureau of the Census.*—Probably the best known of all Federal microdata bases are the public-use samples of basic records from the 1960 and 1970 censuses of population and housing. From the first release in 1963, these samples have provided nearly the full richness of detail about households derivable from the decennial censuses: age, education, income, occupation, etc., of each family member along with characteristics of the family's housing. The sample originally released in 1963 had little geographic information and the sampling fraction was only 0.1 percent of all U.S. households. As a result of the public acceptance and demonstrated utility of that microdata product, public-use samples from the 1970 census were created with a larger sampling fraction (one-percent) and more specific geographic information (areas as small as 250,000 population were identified). A total of six mutually exclusive one-percent samples were made available—taken together, six percent of the national population. These files are available for purchase by anyone and use is not restricted.

Fairly comparable in content and structure to the census public-use samples are the Annual Demographic Files (ADF) generated each year from the March supplement to the Current Population Survey (CPS). A special provision must be added to the aforementioned disclosure rule since the CPS is an area sample and maps are available which define what areas are included in the first-stage sample. The minimum population criterion becomes 250,000 population within sampled primary sampling units in the area to be identified. For example, since central city, other metropolitan and nonmetropolitan components of the population have been identified on the ADF through 1976, a State with even several million total population was not identifiable if there were less than 250,000 people in sampled nonmetropolitan 'counties. (Beginning with the 1977 ADF, all States will be identified, but with central city and metropolitan residence codes suppressed where necessary—(see page 38). There are no restrictions on use of Annual Demographic Files. Files from a number of other household surveys are also released in a similar manner.

b. *Social Security Administration.*—The Social Security Administration (SSA) makes available from its Continuous Work History Sample system the Longitudinal Employee-Employer Data (LEED) File, containing records for one percent of all employees covered by the Social Security System. For every individual in the file there is age, race, and sex information and a record for each employer in each year since 1957. The employer records indicate the industry, State, county, taxable wages and estimated total wages for the year. Scrambled social security numbers for employees are provided only to users who will be updating the sample with data for subsequent years. Purchasers must enter into a written agreement with SSA specifying the purposes for which the file may be used, prohibiting further dissemination without SSA authorization, and specifically precluding any attempt to identify specific individuals or establishments or to match individual records with information in other files on specific individuals. Annual and quarterly files from the system are also available under the same conditions.

SSA also releases microdata files for general public use, i.e., without any restrictions, from several different sources, including the Longitudinal Retirement History Survey, various surveys of disabled persons, the Survey of the Low-Income Aged and Disabled, and certain match studies using data from the Current Population Survey, IRS and SSA. These files are all based on relatively small samples (less than one-percent of the population) and carry only limited geographic information. Unusual values of variables or combinations of variables are suppressed prior to release of the files.

c. *National Center for Health Statistics.*—The National Center for Health Statistics (NCHS) releases public-use microdata tapes from many of its surveys and statistical programs. These includes tapes from the Health Interview Survey, the Health and Nutrition Examination Surveys, the National Ambulatory Medical Care Survey, the Hospital Discharge Survey, health manpower and health facility inventories, the inventory of family planning service sites, vital statistics for the Nation (natality, mortality, marriage, and divorce), and the national natality and mortality followback surveys. These public-use tapes are reported in a catalog published annually (NCHS, 1976).

One NCHS microdata file quite unlike the examples from other agencies is the file on natality, a 50-percent sample of records from the NCHS birth registration system (100-percent for some States in 1972 and 1973). No other Federal microdata file released exhausts a universe or comes that close. Records on the natality file include the age, race and education of the father and mother, the State and county of residence of the mother, the birth date, legitimacy (if recorded) and several characteristics of the mother's previous childbearing history. Purchasers of NCHS microdata sign a simple statement that the file will be used solely for statistical research or reporting purposes.

d. *National Center for Education Statistics.*—The National Center for Education Statistics has available microdata tapes with information gathered from 22,532 graduates of the high school class of 1972, a probability sample made up of approximately 0.7 percent of the National high school class for that year. Information was collected beginning in the spring of 1972, with followup surveys in October 1974, for the National Longitudinal Study of the High School Class of 1972. School record information, such as grade point average, class rank, and area of study are included along with test results and student-provided information on family background, attitudes, and plans for the future. Periodic follow-ups provide information on activity status and changes in attitudes and plans for the future. Geographic information specifies regions and type of community (e.g. rural, suburb, etc.). These files are available for purchase by anyone, and use is not restricted.

e. *Internal Revenue Service.*—The Internal Revenue Service releases two samples of unidentified individual income tax returns, with 150 data items from each return, for tabulation purposes and to allow simulation of the revenue impact of tax law changes. The Tax Revenue Model for National Estimates, with no geographic information, is available for purchase and unrestricted use. Less than 0.2 percent of all returns are included in that file, although the sampling fraction varies among the classes of taxpayers. The Tax Model for State Estimates, including about 0.3 percent of all returns identified to the State level, is available to State tax agencies for tax administration purposes and, once certainty strata are deleted, it is also made available to the public.

B. Evaluation of the Problem

While microdata are made available so that tabulations or other summarizations can be made, it is the possible scrutiny of individual records that causes concern for the violation of confidentiality. While we are confining our consideration to microdata files with no positive identifiers (e.g., name, address, or social security number) a combination of data elements, such as geographic location, age, race, and occupation, if sufficiently detailed, could identify an individual if known by the investigator in advance. Other information on the microdata record so identified would then be disclosed about the individual, e.g., income, marital history, educational attainment, etc.

This section deals with the likelihood of such disclosure and with the bases for determining, in particular cases, whether or not the risks of disclosure are acceptable.

1. Factors Bearing on the Likelihood of Disclosure

a. *Sample size or fraction of the universe.*—If an investigator were searching for a particular individual in a microdata file, his probability of success would be no greater than the chances that a randomly selected individual's record is present in the file, assuming of course that the investigator had no external way of knowing whether or not the individual was selected into the sample. For instance, in a one-percent sample the chances are 99-to-1 against a particular individual having a record in the file.

In stratified samples the likelihood of selection into the sample may vary from stratum to stratum. Further, in multi-stage samples it may be possible for an outsider to determine that some counties but not others were subject to sampling beyond the first stage. It would then be the sampling fraction within the county that would be relevant, rather than the average or overall sampling rate.

b. *Uniqueness.*—The term uniqueness is used here to characterize the situation where an individual can be distinguished from all other members *in a population* in terms of information available on microdata records. The existence of uniqueness is determined by the size of the population and the degree to which it is segmented by geographic information, and the number and detail of characteristics provided for each unit in the data base.

(1) *Geographic information:* The smaller the population, the more easily an individual can be unique; the larger the population the more likely that his or her set of characteristics is duplicated elsewhere. (Also, the larger the population the more costly would be any linkage attempt.)

Size of the population, or of the smallest segment that can be readily identified, can be varied most directly by varying the amount of geographic information supplied on a microdata file.

Geographic information can be in terms of specific areas (e.g., the State of Maryland) or in terms of type of areas (e.g., size of place or rural) or both. Multiple geographic identifiers in combination may identify a small area, e.g., the rural part of an SMSA, or a small part of an SMSA crossing a State line.

Extraneous sources may also provide information about the location of the respondent: knowledge that only certain areas were surveyed or subject to final stage sampling; sequence of records in the file where they have not been scrambled; the existence of more than one version of a file with different sets of geography identified; and neighborhood, county or PSU summary characteristics if present and matchable to an external source.

(2) *Characteristics of the respondent:* In general, it can be said that the greater the number and detail of characteristics reported about an individual the more likely it is that the individual's representation in the file would be different from that of any other individual in the population. Just 10 characteristics with four categories each create over a million possibilities (4^{10}), and when one considers that some data items may have 100 or more potential categories (e.g., age, occupation, industry, income, place of birth) the number of possibilities become astronomical in a file with a large number of characteristics. Many characteristics are, however, likely to be correlated with one another, thus reducing the degree to which an additional item creates additional unique records. For a given subject the number of categories does not entirely account for its potential in an identification process. Some identify especially small populations, e.g., country of birth of the foreign born.

It might then seem reasonable to designate a minimum category population, e.g., to collapse country of birth categories with less than 50 cases in the file. This technique, however, appears inadequate. While there may be many Russian-born persons sampled, only one may be black, or only one may live in a particular identified area. More importantly, uniqueness in the sample is not the critical factor, for there may be a hundred such individuals in the population with no possibility of discriminating among them. Uniqueness in the population is the real question, and this cannot be determined without a census or administrative file exhausting the population or an identifiable subset thereof (e.g., a file of all doctors). Precluding uniqueness in the sample would be a very conservative approach to avoiding disclosure.

Some public-use microdata files provide characteristics for all or at least multiple members of a household. The association of the characteristics of household members greatly increases the potential for unique combinations (e.g., a 66-year-old judge married to a 23-year-old actress).

c. *Recognizability.*—The term recognizability is used here to refer to the likelihood that an investigator could accurately associate unique records in the sample with particular individuals in the population and thereby gain additional information about them. A record in the sample may be unique, but if it cannot be linked with a specific person then disclosure cannot occur.

Three factors affecting recognizability are discussed: the existence of a population register, "noise" in the microdata file, and time lag or the degree to which the microdata information has become out-of-date for an individual.

(1) *Population registers:* A population register is defined here to be a list of persons or households with specific identification, names or addresses, which also systematically contains information which coincides with data on public-use microdata records. Except for Census Bureau, Social Security Administration and Internal Revenue Service records, none of which are available to the public, we know of no registers which systematically cover most of the U.S. population. But neither nationwide coverage nor coverage of all segments of the population is required to make a population register useable for matching purposes.

Reasonable coverage of a defined subpopulation, along with a number of reliable matching characteristics, may suffice. A register of some groups like Black architects, American Indians, high public officials, or birth records is not at all improbable. The existence of rather extensive registers of business establishments in the hands of governmental units, trade associations and firms like Dun and Bradstreet has virtually ruled out the possibility of releasing microdata files about businesses for statistical purposes.

The point is, of course, to be able to discriminate among the units on the register for the one which matches a public-use microdata record, and this requires inclusion on the register of stable and reliable matching characteristics. Among the characteristics most likely to reside in a population register file, date of birth and State or country of birth would seem to be the most reliable, regardless of time or circumstances of data collection. Veteran status, period of military service, and years of school completed would also be consistently reported in different files. Place of residence, family composition, occupation and industry are excellent differentiating characteristics, but since they are subject to change over time, it is more important that the register have been compiled near the time of the census, survey, or administrative action producing the microdata. Further, occupation and industry may be subject to different interpretations or coding errors. Date of first marriage, race and Spanish surname would also be helpful where present. The items mentioned here are the kind of items present in the *Congressional Directory* and in *Who's Who in America,* and need not be associated with dossiers of an investigative agency. Housing characteristics, income, and other characteristics are less likely to be available except by the investigator's own knowledge or inference, and thus may serve to confirm a match while not being too useful in the matching itself.

Neither the *Congressional Directory* nor any of the Who's Who publications is computerized, though the information is presented in a systematic way. Welfare agencies and credit bureaus might have information useable for matching in computerized form, although access to these files is assumed to be restricted. It is also assumed that city directories, voter registration lists, or the records of motor vehicle agencies, tax assessors or real estate agencies would not contain a broad enough set of characteristics for matching, at least with the microdata files we have examined. There should be no doubt, however, that any new file considered for availability in microdata form should be reviewed for its correspondence to various existing population registers.

(2) *"Noise" in the data:* This section deals with inaccuracy and "noise" (random error) in public-use microdata as it affects disclosure potential. Noise may be of two types: that which enters unintentionally during the data collection and processing

and is normally regarded as undesirable, and that which is introduced intentionally during creation of a public-use file so as to reduce disclosure potential. Whatever the source of error—respondent mistake, intentional misrepresentation, coding, transcription, or processing error—unreliability in the microrecord has a direct effect on matchability of the data to a referent in the population. The effect is more severe to attempted identification through matching than it is to the more appropriate statistical uses because there is no chance for compensating errors to average out or to appear small in perspective. If a user were to allow some uncertainty in the matching (e.g., match on the basis of five characteristics even if a sixth was not consistent with the match) the user could not be sure the match was correct. (By comparison, "uncertain matching" is a useful technique when geocoding address files, i.e. when matching address files with geographic base files (GBF's). In this case the GBF presumably has exhaustive coverage of street or city names, and if an address fails to match any record in the GBF it is assumed that there is a mistake in the address (e.g., a misspelled street name). Sophisticated procedures have been developed to match the address to the most similar street record in the GBF using procedures allowing a predetermined amount of uncertainty.)

If unintended error or unreliability helps reduce disclosure potential, then intentional noise added to a microdata file could be still more effective, particularly in touching all records rather than just some. Doing so without damaging the usefulness of the file for statistical purposes is the problem.

(3) *Time lag:* There is inevitably some time lag between the date of data collection or reference date and the date the microdata become available, usually at least several months and sometimes several years. As the data become less current they become less useful for many statistical purposes, but they may also become less potentially dangerous to confidentiality. First, the user will have greater difficulty in reconstructing a given individual's characteristics as of the reference date. Second, whatever possible gain the user might expect from the match will presumably be less. Welfare agencies and credit bureaus might have the best files for matching purposes, but the fact that the linked microdata may be one or more years out of date should reduce the utility of the match substantially. A microdata file could be withheld from public use for a number of months or years to reduce its disclosure potential, or "old" files could be released with less stringent protection than contemporary files.

d. *Hypothesized relationships among the various factors in two types of attempts to penetrate disclosure safeguards.*—In examining the relative impact of the various factors on disclosure potential, it is useful to hypothesize how an investigator might go about trying to identify microdata records. There appear to be two different broad types of potential disclosure situations, and they are affected by the various factors in differing degrees. The first scenario is where the investigator searches the file for a specific individual, using certain characteristics of which he or she is already aware. The second is where the investigator is just "fishing" for a record with a set of characteristics he or she recognizes.

(1) *Searching for a specific individual:* This type of use is the more volatile. If a public-use microdata file were to prove useful for private investigatory purposes, the breach of confidentiality would be extremely serious. The most obvious factor working against misuse of this type is the sample size. In the Annual Demographic File the chances against finding a particular subject would be about 1,399 out of 1,400 even if the best matching variables were known. (If the investigators knew primary sampling unit definitions, the chances might reduce to about 199 out of 200 or so for certain geographic areas.) Even considering the simultaneous use of all six 1970 census public-use samples, and under hypothetically perfect matching conditions, the 94-percent probability of failure should discourage the investigator. Only where there is an extremely large number of subjects for whom excellent matching data are available, and under conditions where success in only a few cases will suffice, could the file seem to be of any use. The existence of some sort of population register or inventory would be almost a necessity for investigatory use. With a population register, any reliable and well differentiated (many categories) matching characteristics will serve the matching process; i.e., geographic information is no more important than other equally detailed matching variables.

It is also true that any substantial noise or inaccuracy in the data would preclude an exact match rather effectively. In fact the introduction of noise would seem to be the best single answer to disclosure if it were not for the resulting inhibition of statistical use.

(2) *"Fishing expedition":* In this type of disclosure situation the investigator is not searching

for a particular individual, but is just "fishing" for a record with a set of characteristics he or she recognizes. A "fishing expedition" would probably involve intense study of individual sample records with certain salient characteristics, such as individuals with high incomes or unusual occupations. "Success" in such an effort does not immediately seem to be very serious, since there is presumably no profitable purpose to be served by such an investigation. This type of investigation might, however, be undertaken in an attempt to discredit the issuing agency or the practice of releasing microdata.

Since one is not starting with a specific set of target individuals, the low probability of their inclusion in the sample is not a problem to the investigator. The investigator selects certain unusual and highly noticeable characteristics, then extracts corresponding records from the sample. The task then is to recognize known households or individuals among the extracted records. A population register would be useful here, especially one exhaustive of a particular population. In the absence of a population register, geographic information on a file is very important since it may be the most specific matching characteristic known to the investigator. Among subject items, any characteristics which the investigator may have observed assist in the match. Number of characteristics reported is important since the matching will depend on some sort of pattern recognition.

Minor aberrations introduced into the data may not inhibit the match if they do not disturb the general pattern, quite unlike the situation with a population register where a minor discrepancy might defeat the match. Compared to searching for a specific individual, the technical requirements for a fishing expedition are relatively modest.

2. Acceptability of the Disclosure Risk

As was noted in Section A, certain types of microdata can be released without concern for disclosure because they are part of the public record. In other cases disclosure is prohibited by law or by administrative regulations.

a. *Potential harm to the respondent.*—If a person were identified from characteristics in a microdata file, and if that file contained further characteristics not already known by the investigator, then disclosure would occur. Whether harm to the respondent would follow from that disclosure, beyond the invasion of privacy, would depend on whether the further information was of an embarrassing nature or could lead to an undesirable action toward the respondent. Certain financial data, or data dealing with illegal or socially undesirable behavior, for example, could, if disclosed, lead to negative consequences. Other items, such as age or education, might lead to harm if disclosed only in certain unusual circumstances, such as if the data contradicted an application for benefits based on age.

The potential for negative consequences to the respondent decreases for most items as the data grow older or out-of-date.

b. *Potential harm to the agency.*—Relative to summary data, the release of microdata has a higher potential for public misunderstanding. Thus, a disclosure with no particularly harmful consequences to the respondent might, if highly publicized, impair an agency's ability to collect data from an increasingly distrustful public. (Even without an actual disclosure, adverse notice in the press, alleging impropriety in microdata release, could have the same effect.)

c. *Resources available to the misuser.*—Misuse of most microdata is assumed to require large amounts of resources. As computer applications go, record matching is relatively expensive. Critics of conservative microdata policies have frequently pointed to this high cost in conjunction with the assumed low payoff for microdata identification. Nonetheless, as pointed out in B.1.d., resource requirements for investigative use are high, but those of a fishing expendition are relatively modest.

C. Disclosure Prevention Techniques for Public-Use Microdata Files

1. General Tradeoffs

From the foregoing it should be apparent that a number of factors impact on disclosure potential, and also that no one of them alone can be so restricted as to prevent disclosure by itself. A file which exhausts a universe, or comes close, presents considerable disclosure potential if it contains any unique records. Geographic information must be restricted beyond the point where an individual user could be familiar with a significant proportion of the universe. but whether that point comes at 25,000, 250,000 or 1 million will depend on the detail in the file and other restrictions imposed. The Census Bureau has imposed a 250,000 minimum population criterion across the board. but that is in the context that the Bureau normally provides data files with highly detailed subject matter (e.g., single years of age, detailed occupation). No formula has been worked out adequately representing the tradeoffs between level of geographic identification, detail of individual subject items, and sample size.

2. Elimination of Categories Identifying Small Salient Groups

No single data category should be so detailed as to identify a small and easily identifiable group. For Indians, tribal affiliation was collected in the 1970 census but excluded from public-use samples because most tribes were quite small and in many cases could be readily located. Detail on type of institution or other group quarters was also necessarily limited so that a single institution of a given type would not be isolated within an identified area. Providing income groupings so that persons with very high incomes can not be separately identified is a more generalized approach to insuring that corporate executives and other highly recognizable individuals not be so easily distinguished from the rest of the population. A common upper limit for detailed income categories is $50,000 per year, although inflation may soon make a somewhat higher cutoff appropriate.

3. Allowing No Unique Cases

It has been proposed (Fellegi, 1972) that microdata files can be made disclosure-free by making sure that there are no unique records in the file, i.e., that every set of characteristics is replicated at least once. There is little doubt that this standard would prevent disclosure since any match attempt would never result in only a single qualifying individual. This is, however, an unrealistic standard for a file with many data items, since the number of possible combinations would be astronomically high even if all the variables were binary—when in fact relatively few of those data items would be involved in any conceivable match attempt.

That procedure does have some relevance when a particular population register is recognized as threatening the confidentiality of a microdata file, for example, a driver's license file with date of birth, state of birth, sex, and marital status. If a four-dimensional cross tabulation of the microdata within the area to be identified had any cells with only one case, categories could be collapsed or areas redefined until that no longer occurred. If more than one population register existed then the resulting microdata could be subjected to additional cross tabulation. This solution should be recognized as being conservative since it is uniqueness in the population, rather than in the microdata file, which assures matchability. Thus, if possible, the multi-dimensional search for the unique case should be performed in the population register file, rather than in the microdata.

4. Introduction of "Noise" into the Data

Perhaps the simplest method of introducing noise into existing microdata is to add or subtract small amounts at random to values of continuous or interval variables. This could be done to all records or to only as many records as needed to create sufficient uncertainty. The one existing application of noise to a Federal microdata file is of this type: small additive random disturbances have recently been introduced into earnings data from SSA's Continuous Work History Sample.

Clayton and Poole (1976) discuss several techniques for noise introduction, adapted from the recent literature on randomized response. These include multiplicative as well as additive models; also "unrelated question" models in which, with a given probability, the item in question has either the value of the sensitive characteristics or the value of an unrelated characteristic, the distribution of the latter being known. The authors present their research on the impact of various additive and multiplicative models, with varying parameters, on certain measures a user might want to derive. Unfortunately, their study deals only with univariate applications, when, in fact, multivariate analyses are more typical of public-use microdata uses. If noise were introduced into data on age, for example, the user's concern is not just that age distributions can be faithfully reproduced, but that the noise does not distort sensitive relationships, such as between age and educational progress where one is attempting to study the cohorts of students ahead of or behind "normal" progress defined by specific age-grade relationships.

Another method of introducing noise is to match households on the basis of race, age of head, family type and family size; then to interchange certain blocks of other characteristics within the matched pairs or groups. This would leave the distribution of any one variable unchanged, and would preserve relationships among variables in the same block and with matching variables. At the same time, relationships among other variables would be distorted.

Further research is needed into just what kinds of disturbances can be made with minimum statistical impact. Error introduction offers at least the possibility of making available for public use files which must otherwise be restricted, or of adding other useful characteristics, such as more specificity of geographic area, to existing types of files.

The variables to which noise is added should be the ones most likely to be found in population register files.

5. Removal of Well-Known Individuals from the File

If disclosure potential lies primarily with a few people with unusual characteristics, their removal from the file is at least worth considering, rather than eliminating some of the information about all of the population. If more than a handful of such individuals is involved there must be concern about bias resulting from their removal. The originating agency could prepare summary statistics about the individuals removed. Such a procedure should not be relied on to the exclusion of other techniques, since the existence of a large population register would make many people recognizable in a detailed file.

6. Release of Customized Files

Almost any statistical use could probably be accommodated if a microdata file were designed only for that use. For example, the Census Bureau has received requests for customized versions of the Annual Demographic File from CPS identifying a different geographic scheme than that routinely available. Frequently, the requester expresses willingness to do without half or more of the items on the file. Taken alone that request might meet Census Bureau criteria, but since another version of the file is already available the new request is disallowed. The new file could be matched with the old file on a case by case basis, achieving the identification of the intersection of the two geographic schemes.

Census public-use samples for 1970 allowed three alternative geographic schemes by tripling the number of sample cases drawn from the census data base. Cases included in each type of public-use sample were not the same as cases in any other type and could not be matched. When it became necessary to produce another public-use sample for a special purpose, it was possible to draw still another sample from the data base. This luxury of offering multiple microdata bases from a single source is only practical with a census or set of administrative records containing far more individuals than would be needed on a single public-use file. The Census Bureau has not, for instance, seriously considered subdividing the Annual Demographic File into two half-samples, since having the full sample size is deemed more important than offering geographic options.

Customized files are feasible only in contexts where there is no violation of standards if the information in all available files were combined, or where it can be guaranteed that there can be no matching between two files with the same cases. In the latter case, the customized files would not be public-use files. Release of files for restricted use is discussed in the next section.

D. Disclosure Prevention Through Restrictions on Use

1. Alternatives Where Public-Use Microdata Are Not Satisfactory

a. *Special tabulations by the originating agency.*—The researcher whose needs are not met by public-use microdata normally has the alternative of paying the source agency to make special tabulations of the source file, to give him the same tabulations he would have created himself. Researchers frequently do not find this type of arrangement satisfactory. Agencies are rarely able to maintain enough staff so that special tabulations can be handled without delay. The researcher is expected to reimburse the agency for programming and computer time and for administrative overheads, usually at rates above levels he would pay at his own institution. The process of getting results, deciding on revised specifications and repeating the process perhaps more than once becomes cumbersome when working through layers of intermediaries. Also some of the desired statistical operations, e.g., the transfer-income model, are so sophisticated that it often becomes impractical for the source agency to perform the task.

b. *Microdata available for restricted use.*—It would seem reasonable that microdata which do not meet the requirements for public use should be usable outside the originating agency if it were possible to require the user to observe the same restrictions as the originating agency observes to guarantee confidentiality.

2. Contractual/Administrative Requirements on the Restricted User

Restricted-use arrangements would be designed to contractually bind the user to the same precautions taken by the originating agency. The following are examples of conditions which might be applied in the release of microdata for restricted use:

a. The activity must be in accordance with the laws governing the programs of the respective agencies.

b. There must be a legitimate and important research purpose to be served by the process.

c. The receiving agency must be strictly and legally accountable to the providing agency for its security program.

d. The receiving agency must demonstrate that it has adequate security protections.

e. The microdata would contain no individual identifiers nor typically contain data which would be easily associated with an individual.

f. The receiving agency would not and could not be required to turn the data over to any third party, even under subpoena or a Freedom of Information Act request.

g. The providing agency would have opportunity to review any publication of information from the data to insure that no potential disclosures are published.

h. At the conclusion of the project, and no later than some specified date, the receiving agency would either return or destroy all of the microdata involved.

i. Significant sanctions or penalties for improper disclosure would apply.

Certain of the foregoing conditions would probably not be possible for most agencies without changes to existing legislation, as in the application of criminal penalties for improper disclosure (item i) or in guaranteeing immunity from legal process (item f).

The Privacy Protection Study Commission, in its final report, has recommended that such releases be allowed under a similar set of conditions, and has called for legislative action to establish these conditions (Privacy Protection Study Commission, 1977, Chapter 15.)

3. Agency Experience with Use-Restricting Agreements

a. *Bureau of the Census.*—Purchasers of the 1-in-1000/1-in-10,000 1960 census public use samples issued in 1963 signed an agreement (1) prohibiting any dissemination of the samples to a third party without written authorization from the Census Bureau, (2) requiring that any publications incorporating data from the samples contain a standard disclaimer paragraph, and (3) requiring that the Census Bureau be provided a copy of any publication containing data derived from those data files. Purchasers were reminded of these obligations in a supplement to the file documentation issued in 1964. 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.

From this experience it was apparent that the sample purchasers either did not take their signed agreement seriously, forgot it after a period of time, or were not able to control handling of the file at their institutions. In a few cases the agreement had been signed by a university purchasing agent and was unknown to the actual users. This experience suggests the necessity of more complete arrangements with purchasers of restricted-use files, including periodic followup, and denying access to researchers who are not able to control completely the handling of the data files in question within their institutions.

b. *Other agencies.*—Neither the Social Security Administration nor the National Center for Health Statistics has detected any violations of their use-restricting agreements, although it should also be said that neither agency has felt it necessary to undertake systematic monitoring to detect potential abuses.

4. Relationship of Computer Security to Use Restriction

Administrative restrictions on how a file is used cannot be effective without appropriate security in the computer system in which restricted data are used or stored.

At the simplest level restricted files on tapes or other storage devices must be protected from theft or unauthorized copying. The computer operating system must be capable of detecting and preventing unauthorized access. Files or parts of files may be further protected by passwords and encryption algorithms. Appendix B discusses the various aspects of computer security and cites some of the current literature on the topic.

CHAPTER V

The Question of Balance: Protection of Individuals vs. Public Needs for Information

A. Introduction

The establishment of suitable disclosure-avoidance policies requires a balancing of conflicting objectives. The situation is somewhat analogous to statistical hypothesis testing or quality control, where findings or decisions are subject to errors of two kinds. With respect to any specific disclosure-avoidance procedure, we might define—

Errors of the first kind, i.e. publication or release of information that can be associated with specific individuals (or other statistical units), possibly resulting in harm or embarrassment to those individuals.

Errors of the second kind, i.e. suppression or withholding, for the purpose of avoiding disclosure, of statistical information which if released could have benefited society in significant ways.

It is unlikely that policies can be adopted which will guarantee complete elimination of either type of error without increasing the other type of error to an unacceptable level. Compromise is unavoidable.

Thus, it is necessary to introduce an additional concept subordinate to the broad definition of statistical disclosure presented in Chapter II, namely that such disclosure may be *acceptable* or *unacceptable*, depending on the particular circumstances in each case. We cannot provide a single definition of these terms which will cover all situations. In the last analysis, the selection of disclosure-avoidance techniques is a matter of public policy, representing an acceptable balancing of conflicting objectives, and cannot be resolved by this Subcommittee. However, the Subcommittee felt that it could make a contribution to informed discussion of this question, first by reviewing what government statisticians have had to say about the issue of balance, and second by searching for and reporting on instances where individuals or groups have expressed dissatisfaction with specific disclosure avoidance policies, either as not being sufficiently protective of individuals, or as resulting in too much withholding of needed statistical information.

B. Comments in the Literature

A study of the literature makes it clear that producers of government statistics and others who have studied the question of disclosure are increasingly aware that there is no such thing as absolute protection from statistical disclosure, and that the operational problem is one of striking a suitable balance between the two kinds of "errors" mentioned earlier.

In a paper presented before the International Statistical Institute (Barabba and Kaplan, 1975), a former director of the U.S. Census Bureau and one of his colleagues reviewed in some detail the policies and procedures of the Bureau of the Census for avoiding disclosure. Their conclusions were as follows:

> The U.S. Census Bureau has a long and continuing history of protecting the confidentiality of information it receives from individual people and businesses. 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.
>
> Therefore, the Bureau is convinced that both the fact and perception of its protective techniques must be unambiguous. In some contradiction to this aim is society's growing need for information, especially on a small area basis. The protective techniques should, therefore, not be so overwhelming as to markedly damage the usefulness of the data. A balance must be struck. Developing techniques which maximize protection against disclosure while minimizing dis-

ruptions in the data product is, for the U.S. Census Bureau, a high priority task.

Ivan Fellegi, an official of Statistics Canada, in an article "On the Question of Statistical Confidentiality" (Fellegi, 1972), made the following statement concerning the release of microdata files:

> Even though the release of census data for a sample of individuals may, in a rigorous interpretation of the concept, be disclosure it can be argued that the probable pay-off to anyone looking for information about a particular person is sufficiently small, while at the same time the benefit to users of such tapes (and, indirectly, to society) is sufficiently large that the cost-benefit ratio to society is highly favorable. Obviously, pragmatic considerations must be taken into account.

Tore Dalenius (1974) made the following statement concerning the balancing problem faced by producers of statistics:

> The producers have clearly a most subtle task: to strike a reasonable balance between publishing "too much" and thus exposing some objects to the risk of exposure, and publishing "too little," thus depriving users of valuable information. It must be expected that now and then mistakes are made, and it seems obvious that publishing "too much" arouses more and louder criticism than the opposite mistake.

Finally, a very succinct statement of the balancing issue was given by Morris Hansen in a chapter he wrote for the Report of the President's Commission on Federal Statistics (1971):

> It is desirable to have recognized in applying past principles and in developing any new ones, that if *any* statistics are to be published nondisclosure cannot be absolute. Rules for nondisclosure are necessarily based on an interpretation of what is reasonable, and supported by precedents and past experience.

C. Reactions to Agency Policies and Procedures for Disclosure Avoidance

To place Federal agency policies and procedures for disclosure avoidance in broader perspective, the Subcommittee sought to ascertain their impact, both within and outside government circles. First, we consider the impact upon data subjects. The information at our disposal suggests marked differences between individuals and organizations as data subjects. Hence, they are discussed separately. Thereafter the impact of disclosure-avoidance policies upon data users is traced. The discussion concludes with a brief note about the portrayal of agency disclosure-avoidance practices by commentators on the subject.

There is a general lack of documented information about reaction to agency practices. Accordingly, we report what little evidence was available at the time this report was written. Further information bearing upon these issues is welcome.

1. Impact on Individual Data Subjects

The chief concern about individual data subjects is the possibility that a data release from a Federal agency could permit disclosure that might cause actual harm to an individual. Accordingly the Subcommittee sought out evidence of harm. *The Statistical Reporter* (No. 77–4, January 1977: pp. 137-138) included a request for "information about any harm which may have befallen an individual . . . as a result of statistical disclosure." No information has been received in response. A similar appeal for information was addressed to Carole Parsons, Executive Director, Privacy Protection Study Commission, again with negative results. Informally, members of the Subcommittee canvassed their colleagues for relevant information, again to no avail.

The Subcommittee found only one class of allegation of harm from statistical disclosure. Several individuals have complained that the release of population census summary data by zip-code area has contributed to their increasing receipt of junk mail. Such allegations do not imply that any information about particular individuals has been released—merely that a particular kind of group disclosure encouraged junk mailing.

Repeated attempts have not succeeded in locating any other instance in which an individual data subject alleged that he or she had been (or might be) harmed in any way by statistical disclosure. While further investigation into the matter may uncover isolated instances of harm, there is no indication that any agency releasing statistical data is harming data respondents through improper data-release practices.

There is a second and somewhat related line of inquiry being undertaken with regard to the level of comprehension and satisfaction of individual data subjects regarding agency policies and procedures for disclosure avoidance. In survey research it is an article of faith that strong confidentiality measures are needed to warrant the public trust and minimize the refusal rate. However, the importance ascribed to the public trust stands in marked contrast with the amount of evidence available on it. Federal statistical agencies do not routinely collect information on data subjects' views of disclosure-avoidance measures. However, there are now underway several Federally sponsored studies to close this gap in knowledge.

One study of the impact of confidentiality pledges upon data subjects was conducted by Eleanor Singer of the National Opinion Research Center, New York office, with funds from the National Science Foundation. The experimental (factorial) design involved the manipulation of several independent variables including the assurance of confidentiality, which was varied to include no mention, an absolute guarantee, and a qualified guarantee—"except as required by law."[1] Dependent variables included: (1) response rate to the interview as a whole; (2) response rates to different types of questions within the interview (e.g., more or less threatening, factual vs. opinion); and (3) quality of response. At the end of the interview, respondents completed a self-administered questionnaire asking for their reactions to the interview and for their recollections of what the interviewer had said about confidentiality, voluntary participation, sponsorship, etc. In a follow-up letter, respondents were informed that the assurance of confidentiality as well as other elements of the introduction had been varied among respondents in order "to know the best way to describe a study so that respondents have enough information to decide whether or not to participate in it." The letter added "*All* information will, of course, be treated as confidential and the data will be presented only in aggregate form."[2] Singer (1977) gives a preliminary report of the findings.

In order to conduct studies with similar purposes, the Committee on National Statistics of the National Academy of Sciences has formed a Panel on Privacy and Confidentiality as Factors in Survey Responses with funding from the Bureau of the Census. One study being undertaken by the Panel, under the direction of Edwin Goldfield, examines how the confidentiality pledge affects responses, by varying the length of time in which answers would purportedly be kept confidential by the Census Bureau among four alternatives ranging from everlasting confidentiality (i.e. unlimited duration) to no confidentiality (i.e. the collected information about identified individuals could be immediately available to other agencies and the public). To a fifth segment of the sample, no statement of confidentiality is given.

At the conclusion of the interview the respondent is asked to recall the terms of confidentiality, if any, that were stated at the outset. Finally, respondents are handed a letter that thanks them for participating in the experiment and assures them that in fact their answers will be kept confidential for as long as the questionnaires remain in existence.

Both the response rate and the quality of data will be examined according to the degree of confidentiality promised. The accuracy of recall of the confidentiality pledge will serve as yet another gauge of respondent concern with the confidentiality issue. Where possible, responses will be validated against Census Bureau records.

A second study being conducted by the Panel is an opinion survey of 1,500 households regarding their perceptions and attitudes toward confidentiality, privacy, and other factors thought to influence survey response. Among the issues examined is that of intrusiveness, i.e., do respondents feel that the Federal government collects more information about individuals than it needs? Are questions pertaining to age, sex, race, education, income, social security number, etc., proper topics for government inquiry, as far as the respondents are concerned? Throughout the interview respondents are asked to distinguish among the Federal government, State or local governments, universities and private companies as takers of surveys. To study the effects of study sponsorship, data collection is being conducted by both the Bureau of the Census and the Survey Research Center of the University of Michigan, each covering a random half of the sample.

Goldfield *et al* (1977) report preliminary results for both studies.

[1] Interviewers were supplied with different sets of responses, according to the level of confidentiality promised, in the event a respondent queried the interviewer about confidentiality.

Where interviewers were not to mention confidentiality, if respondents asked whether their replies were to be kept confidential, interviewers were instructed to respond, "I really don't know. I know that respondents' names are never used in reports." They were explicitly instructed not to promise confidentiality.

Under absolute confidentiality conditions, interviewers were instructed to respond, "Yes, your answers will be kept confidential." If respondents asked about procedures for keeping replies confidential, interviewers were instructed to say "Well, no one is ever identified by name in reports. Responses are used for statistical purposes only."

Under qualified confidentiality conditions, interviewers were instructed to respond, "We will do our best to protect the confidentiality of your answers." If the respondent asked "How do you protect the confidentiality of answers?" interviewers were instructed to respond, "No one is ever identified by name in reports, but if names are subpoenaed, NORC would have to obey a court ruling."

[2] Study materials supplied by E. Singer.

2. Organizations as Data Subjects

Organizations have registered a clear concern with agency disclosure practices. An example of this concern is the litigation against the Line of Business (LOB) Program of the Federal Trade Commission (FTC) by a number of large corporations. The LOB program seeks detailed financial and related data for every line of business in which a given corporation has sales or receipts totalling at least $10,000,000. For this program the FTC has delineated some 261 different manufacturing industry categories as of 1976.

Corporations opposing the LOB program claim the FTC's publication plans amount to statistical disclosure. More specifically, the corporations state that if the FTC publishes LOB totals for each LOB in which there are four or more reporting companies, as planned, it will be possible to examine these totals together with other company, market, and industrial information and determine exact or approximate values of data for individual companies. They presented in court an estimation procedure intended to substantiate their claim, by calculating some matrix elements and narrowing the uncertainty range for the remainder. Involved were mathematical techniques for solving linear equation systems whose variables are subject to additional linear inequality constraints.

The FTC has pledged not to publish any aggregate statistics based upon data for fewer than four firms. Furthermore, it will publish no number which would permit the determination of an aggregate figure for less than four firms. For example, if seven firms filed LOB reports for a particular industry category, aggregated data for that category as a whole could be published. However, it would not then also publish statistics for the four largest reporting companies in that LOB, since that would make it possible to determine the aggregate statistics for the three smallest firms by subtraction. The FTC staff also intends to perform special analyses to ensure that no accidental disclosures result from the publication of aggregates based on four or more firms. To determine what additional tests might be necessary, the Commission invited companies to articulate any special conditions which might facilitate disaggregation. According to the FTC, the responses to this invitation did identify a few special circumstances which posed a threat of disclosing individual company data, even after the "four or more" criterion had been imposed. Responses also included standard techniques for constructing interval estimates of establishment data.

As for the estimation of individual company data, the FTC acknowledged that there are techniques whose application can establish ranges within which an individual firm's data must lie. However, the FTC emphasizes that it does not guarantee that its published report cannot be used to make estimates: the guarantee is that the Commission will not publish aggregate numbers from which it is possible to go beyond estimating the components of the aggregate to knowing their exact values. In short, the FTC publication plans permit approximate (but not exact) disclosure.

The LOB experience suggests that businesses may be more concerned than individual citizens about the possibility of statistical disclosure about themselves. However, the experience of the LOB program should not necessarily be construed as typical of the Federal experience in dealing with data from companies. To mention a contrasting example, the Bureau of the Census has collected and published extensive data from the same companies over the years in its censuses and surveys (e.g., the Census of Manufacturers) without similar complications.

3. Reactions of Data Users

We turn now to discuss the reactions of data users to agency disclosure-avoidance practices. No specific studies could be located concerning the effects on users resulting from the suppression or alteration of information by producers of statistics in order to avoid disclosure. The Subcommittee's information on the subject is anecdotal, based largely on personal experience.

The main difficulty voiced both within and outside the Federal government seems to be the suppression of important data elements pertaining to non-identified individuals. This comes about because data elements which would tend to identify individuals are routinely edited from table shells or stripped from microdata files. For example, the Bureau of the Census removes from public-use microdata files any data element which would identify an individual as living in a particular area with a total population of less than 250,000. Users complain that this results in a loss of data for analytical purposes and sometimes makes it impossible for users to calculate sampling errors for the statistics of interest whenever the information on sampling errors provided by the releasing agency is not adequate for the users' purposes.

a. *Data-loss problem.*—An illustration of the data-loss problem can be taken from materials presented before the Privacy Protection Study Commission. Specifically, for a study of postsecondary school enrollment sponsored by the National Center for Education Statistics (NCES), the Bureau of the Census collected information from a national sample of students about their receipt of financial aid, which schools they were attending, etc. However, the Census Bureau is not willing to transfer to NCES the names of postsecondary educational institutions reported by students, on the grounds that this information would tend to permit respondent identification: its transfer would thus violate the provisions of Title 13. This decision thwarts NCES plans to analyze the data on individuals in combination with the extensive information on postsecondary institutions which NCES collects. At issue are such questions as how student financial aid is distributed among institutions of varying characteristics. While the Bureau of the Census could add the institutional information to its data file, the enlarged file could not then be forwarded to NCES shorn of institutional identifiers, because the institutional characteristics would tend to reveal institutional identity. The alternative of analyzing the data through Census Bureau facilities is cumbersome and slow at best. Mandated by law to "report full and complete statistics on the condition of education in the United States," NCES pointed out to the Privacy Protection Study Commission how the disclosure practices of one Federal agency can limit another statistical unit in the pursuit of its legislated mission.

Additional illustrations are provided in a report which the Subcommittee has received from the Law Enforcement Assistance Administration (LEAA). This detailed statement (LEAA: 1977) tells how the Census Bureau's disclosure-avoidance requirements have, in LEAA's opinion, unduly limited the utility of data collected by the Census Bureau for LEAA. The LEAA feels that the Census Bureau's disclosure-avoidance policies, as applied to the release of tabulations and microdata from the National Crime Survey and the Juvenile Detention and Correctional Facility Census, are so stringent that they have seriously handicapped uses of data from those surveys, especially by users who are interested in particular States and cities. According to the LEAA, detail needed for secondary analysis, including the evaluation of sampling and nonsampling errors in the primary data, could not be made available to LEAA itself or to other potential users. The LEAA also states that Census Bureau practices have recently become more restrictive, so that public-use microdata tapes for the State of California from the National Crime Survey, which had been released for 1973 and 1974, were not released by Census for 1975. The LEAA statement concludes that "The net effect of the current Census Bureau practices, as illustrated by the National Crime Survey and the Juvenile Detention Correctional Facility Census, is to prevent and seriously restrict LEAA's efforts to improve, redesign, and expand the use of these statistical series."

b. *Crosscutting standard geographic areas.*—Disclosure-avoidance techniques are invoked bearing in mind all statistical releases from a given data set, not just the particular release at hand. The rationale, of course, is that this precludes the possibility of piecing together potentially identifiable information from complementary releases. This general approach thwarts requests for data by groups with an interest in information that crosscuts standard categories. Regional commissions illustrate the case, since several regions do not conform to State boundaries, which is the common mode of data release. The Appalachian Regional Commission (ARC), for example, encompasses only one entire state (West Virginia) while overlapping partially with 12 others. Because sectoral employment and income information is published at the State level, ARC finds it impossible to obtain complete data on the region from the Bureau of Economic Analysis (BEA). Recent tabulations furnished by BEA give only range estimates of the total amount of earnings in the region in selected industrial sectors. ARC regards this as insufficient and an example of "the growing tendency of (Federal) government agencies to withhold information from other public agencies." BEA, however, points out that it is only an intermediary using Bureau of Labor Statistics (BLS) data and conforming to BLS disclosure rules.

c. *Changes in disclosure-avoidance techniques.*—A specific instance of an expression of frustration about *changes* in disclosure-avoidance practices came to light in the report of a survey on the timeliness of Federal statistics conducted by the Federal Statistics Users' Conference. A respondent to the survey was quoted in the report as objecting to the disclosure-avoidance procedure for the 1972 Census of Retail Trade:

> In 1967, county totals were rarely deleted for SIC volume totals unless a single organization completely dominated a county. Now, a single nonemployer can cause these county totals to be deleted to avoid disclosure. Someone made a poor decision.

This user purchases special tabulations, by county, from each quinquennial census of retail trade. In standard format for these tabulations, county data are provided separately for establishments with and without employees, and for the two groups combined. In the tabulations from the 1967 census, potential disclosures were avoided by suppressing the detail and showing only the totals for all establishments in a county. In the tabulations from the 1972 census, if there were a potential disclosure in the data about establishments without employees, disclosures were avoided by suppressing the county totals and the data were shown only for establishments with employees. This change, which was made for all census tabulations, and thus was made without consulting the purchaser, affected the comparisons of 1967 and 1972 data in several hundred counties. From the Bureau's perspective, at issue was which of two sets of statistics should be given priority for public release, if both could not be shown.

d. *Changes in methodology.*—Changes in data-collection procedures for a time-series data set sometimes trigger changes in the disclosure-avoidance practices, thus creating discontinuities in the data available. For example, the sample for the Current Population Survey (CPS) was recently enlarged, making it possible in 1977 for the first time to identify all States in the public-use microdata releases. This will satisfy the demands of many users heretofore not able to make full use of the CPS. At the same time, the identification of small States will preclude the identification of central city and metropolitan or nonmetropolitan residence in certain parts of the country, which has been available in the past. This change, as with the Census of Retail Trade example mentioned above, will interrupt time-series analyses or otherwise thwart the interests of some present users, even though it serves a number of new users.

e. *Data-users' options.*—Accommodating to agency procedures, data users sometimes are allowed to establish the priority of available data elements, so as to obtain as much of the needed information as possible. In many instances this involves forfeiting certain kinds of geographic detail in order to obtain the key geographic dimension (e.g., identities of States vs. metropolitan-nonmetropolitan, and within that, the distinctions among central city, non-central city, urban and rural). Reviewing the sampling plan, the collecting agency can inform the data user as to how many geographic units of each type would be identified using alternative priority schemes. The choice of geographic detail is particularly limiting in situations where an agency plans to make many different uses of a data set and the level of detail needed for one purpose preempts the level needed for other purposes.

Flexibility in giving data users a degree of choice is limited by the number and diversity of public-use tapes or tabulations being prepared on a given data set, as well as by the release of data from related sets. The existence of additional information presents the possibility of penetrating the anonymity of information within any one data set. In some cases users may be able to obtain special tabulations or special-purpose microdata files tailored to meet specific needs not met by an agency's original data releases. However, considerations of cost, time, and the complexity of data manipulations frequently limit the utility of this option.

f. *Recommendation by the Census Advisory Committee of the American Statistical Association.*—The Census Advisory Committee of the American Statistical Association (ASA) has discussed disclosure issues in two of its recent meetings. While this Committee is not strictly a user group, its reaction to the Census Bureau presentation at the earlier of the two meetings (ASA Census Advisory Committee, 1975) indicates its concern that user interests be considered in the establishment of disclosure-avoidance policies. The ASA recommendation follows:

> The Committee is impressed and pleased by the continuing efforts to protect confidentiality. The technical developments on minimal masking of cell results needed to protect confidentiality are especially interesting.
>
> Nonetheless, we feel that there is some danger of overreaction to the threat to confidentiality from inspection of tables or sophisticated statistical analysis of Census data. This threat is surely small compared to the threat of direct access to questionnaires, which the Census Bureau has always defended against, even during wartime. In particular, we think the risk of revelation of sensitive demographic information about individuals in small-area tabulations of Census data is smaller by several orders of magnitude. The corresponding risks in economic censuses are probably much greater, but even here we caution against the temptation to set excessive standards of protection that are needed to foil a determined analyst, armed with lots of cleverness, determination, computer funds, and a good knowledge of mathematics.

4. Reactions of Others

Commentators have alleged occasionally that statistical releases violate confidentiality, even where there is no apparent ability to identify a specific individual. To cite an example, in a 1967 article entitled "The Punchcard Snoopers," Phil Hirsch argued that the release of a summarized income distribution for each of six groups of Illinois doctors (general practitioners, internists, and surgeons; inside and outside the Chicago area) violated the 1960 census questionnaire's promise of confidentiality. In fact, that release provided no information about particular doctors. The income information for the three medical categories was derived by the Census Bureau's matching its 25-percent sample records against records for 900 Illinois doctors provided by the American Medical Association (AMA), which resulted in 188 successful matches. Thus, while the data released did reveal information about a sample from an identifiable group, there was no disclosure of information on an identifiable individual.

Several years later, a prominent author on privacy issues picked up on the Hirsch article and, apparently misinterpreting Hirsch, alleged that the Census had made the income data available in a list of the Illinois doctors where "identification of individual doctors was possible" (Miller, 1971: 136).

CHAPTER VI

Findings and Recommendations

A. The Concept of Statistical Disclosure

Findings

Several of the major Federal statistical agencies have developed and applied a variety of *disclosure avoidance techniques* in connection with the release of statistical tabulations and microdata files (files of individual records with identifiers removed). However, it appears that little attention has been given to defining exactly what constitutes disclosure and how to decide which disclosures are acceptable and which are not.

A few statisticians, notably Fellegi (1972), Hansen (1971), and Dalenius (1977) have suggested formal definitions of statistical disclosure. This Subcommittee has adopted the definition proposed by Dalenius as a framework for its discussion and review of disclosure-avoidance techniques. The Dalenius definition is broad in scope. It was not the intention of Dalenius, nor is it ours, to recommend or imply that statistical disclosure so defined should never be permitted to occur. If that position were adopted, the present output of statistical information would be drastically reduced. We have adopted this broad definition because we believe it offers the best basis to

1. Identify all potential disclosures in connection with proposed releases.
2. Decide which of these potential disclosures are *unacceptable*.
3. Use appropriate techniques to prevent unacceptable disclosures.

The formal definition of disclosure adopted by the Subcommittee appears in Chapter II, pp. 7–10. It can be summarized here by saying that disclosure takes place if the release of tabulations or microdata makes it possible to determine the value of some characteristic of an individual[1] more accurately than would otherwise have been possible.

B. Deciding What to Release

Findings

1. Federal statutes and regulations governing the release of statistical information in the form of tabulations and microdata do not and cannot provide a clear basis for deciding in each case what must be done to avoid disclosure. Agencies that address this issue are obliged to strike a balance between the requirement to protect the confidentiality of information about individuals and the need for detailed statistical information and records for public policy purposes.

2. The use of microdata files by social scientists and others has developed rapidly since 1960. Microdata file users are becoming increasingly adept at handling these files and are applying sophisticated analytical techniques to exploit them fully. This development has significantly increased the utility of statistical data bases created by Federal agencies from censuses, surveys and administrative records and promises to do so even more.

3. The Privacy Act provision concerning the "disclosure" of certain microdata files (552a(b)(5)) is ambiguous and has resulted in considerable uncertainty as to the circumstances under which microdata files can be released.

4. The Subcommittee has identified several examples of statistical disclosure which, *in our opinion*, were not acceptable. Some of these involved potential disclosures of salaries or benefit amounts of specific individuals. We also find, however, that most agencies that release statistical information are becoming increasingly sensitive to the disclosure issue, and that they have adopted or are in the process of adopting policies and procedures designed to avoid unacceptable disclosure (see agency statements in Appendix A).

Recommendations

B 1.. All Federal agencies releasing statistical information, whether in tabular or microdata form, should formulate and apply policies and procedures

[1] Except where otherwise specified, the word "individual" as used in this chapter is meant to cover all types of reporting units—natural persons, corporations, partnerships, fiduciaries, etc.

designed to avoid unacceptable disclosures. Because there are wide variations in the content and format of information released, the Subcommittee does not feel that it is feasible to develop a uniform set of rules, applicable to all agencies, for distinguishing acceptable from unacceptable disclosures. In formulating disclosure avoidance policies, agencies should give particular attention to the sensitivity of different data items. Financial data, such as salaries and wages, benefits, and assets, and data on illegal activities and on activities generally considered to be socially sensitive or undesirable require disclosure-avoidance policies that make the risk of statistical disclosure negligible.

Agencies should avoid framing regulations and policies which define unacceptable statistical disclosure in unnecessarily broad or absolute terms. Agencies should apply a test of reasonableness, i.e., releases should be made in such a way that it is reasonably certain that no information about a specific individual will be disclosed in a manner that can harm that individual.

B 2. Special care should be taken to protect individual data when releases are based on complete (as opposed to sample) files and when data are presented for small areas.

B 3. In formulating disclosure-avoidance policies and procedures, agencies should take into account the various kinds of disclosure discussed in Chapters III and IV of this report. Thus, these policies should deal with situations which can lead to unacceptable disclosures, such as:

a. *In tabulations:*
   (1) Empty data cells.
   (2) Cells equal to marginal totals.
   (3) Cells representing a small number of cases.
   (4) Quantity data cells dominated by one or two units.
   (5) Sets of tables from which the above situations can be arrived at by algebraic manipulation.

b. *In microdata files:*
   (1) Files containing data for all members of a defined population.
   (2) Files with detailed geographic information.
   (3) Files with very precise information, such as exact dates of events, or exact amounts of various kinds of income or assets.
   (4) Files containing substantial amounts of information which is likely to be duplicated in external sources containing identifiers.

B 4. With respect to the release of microdata files the Subcommittee believes that

a. There should be *no restrictions or conditions* attached to the release of microdata files when it is reasonably certain that no information for specific individuals will be disclosed as a result. The Subcommittee has referred to files released under these conditions as *public-use files.*

b. Where the test for a public-use microdata file is not met, but it appears that the public interest will be served by releasing microdata files for statistical and research purposes on a restricted basis to specific users, such releases should be permitted when *all* of the following conditions are met.[2]
   (1) The receiving organization has authority and obligation to protect the file against mandatory disclosure equivalent to that of the releasing agency.
   (2) Responsible personnel of the receiving agency are subject to meaningful sanctions for violations of confidentiality provisions.
   (3) The receiving organization agrees to:
      (a) Use the file only for statistical and research purposes.
      (b) Not attempt to identify individual data subjects for any purpose.
      (c) Not release the file to anyone else without authorization from the releasing agency.
      (d) Maintain adequate security to protect the file from inadvertent or unauthorized disclosure.
      (e) Apply agreed-on disclosure-avoidance techniques before releasing tabulations based on the file.
      (f) Destroy or return the file within a specified period of time.

B 5. With respect to the release of tabulations, a distinction between unrestricted (public-use) and restricted releases, similar to that described for microdata files in recommendation B 4, would also be appropriate. Thus, for tabulations for which the risk of statistical disclosure is deemed too great to permit release to the general public, restricted releases might be made under conditions similar to those described in paragraph b of recommendation B 4, substituting "tabulations" for "file" wherever the latter word appears.

[2] The Subcommittee recognizes that some agencies cannot make this kind of restricted release under existing law.

B 6. To insure compliance with its disclosure-avoidance policies and procedures, each agency that releases statistical information should establish appropriate internal clearance procedures. There should be a clear assignment of individual responsibilities for compliance. Staff members responsible for compliance should be encouraged to become familiar with the materials summarized in this report, and to take advantage of relevant training activities (see recommendation C 2).

B 7. In order to guide their disclosure-avoidance policies, agencies should systematically document the consequences of these policies. In particular they should investigate and record:

a. The details of any cases in which data subjects or others allege that statistical disclosure has occurred.

b. Requests for tabulations and microdata files without identifiers that have been denied or only partially met because of agency disclosure-avoidance policies.

B 8. The Office of Federal Statistical Policy and Standards (OFSPS) should encourage agencies that release tabulations and microdata to develop appropriate policies and guidelines for avoiding disclosure, and to review these policies periodically. To the extent feasible, OFSPS should help agencies to obtain technical asistance in the development of disclosure-avoidance techniques. OFSPS should also be prepared to assist and advise agencies in cases where unacceptable disclosures are alleged to have occurred and in cases where potential users, including other Federal agencies, feel that agency disclosure-avoidance policies are unnecessarily restrictive.

C. Disclosure-Avoidance Techniques

Findings

1. In recent years, many different effective techniques for avoiding disclosure have been developed and used. No one technique is ideal for all types of releases.

2. While these techniques have been applied in several instances in the United States and other countries, they are not generally known or accessible to many agency personnel responsible for the release of statistical information. In this report, we have tried to provide a systematic summary description of useful disclosure-avoidance techniques and references to more detailed information.

Recommendations

C 1. This report should be given wide circulation to Federal agencies that release statistical information, whether based on surveys or on program records.

C 2. Based on the material covered in this report, the Office of Federal Statistical Policy and Standards should conduct periodic training seminars for Federal agency personnel who are responsible for developing and applying statistical disclosure-avoidance procedures. These seminars could be organized in much the same way as OMB's recent seminar on presentation of errors in statistical data. Participants would be expected to train and provide technical assistance to appropriate persons in their agencies.

C 3. Disclosure-avoidance procedures should be described, in a *general* way, in connection with publications or other releases of data to which the procedures have been applied. However, such descriptions should not include details whose publication would tend to reduce the degree of protection provided by the particular procedures used.

C 4. To minimize disclosure risks, agencies that release data based on samples should, where feasible, refrain from publishing information that would make it easier for others to determine which individuals were included in the sample. For example, if a sample is based on ending digits of social security numbers, the particular pattern of ending digits used to select the sample should not be published.

D. Effects of Disclosure on Data Subjects and Users

Findings

1. 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. The same statement cannot be made for legal persons (corporations, partnerships, etc.) as data subjects. Several companies included in the Federal Trade Commission's Line of Business Surveys have sought legal relief from mandatory responses, asserting that publication of tabulations as planned by FTC would result in damaging disclosures of individual company data.

2. There have been a number of cases in which

users of data for both natural and legal persons have been unable to obtain the amount of detail desired from tabulations or microdata files because of agency disclosure-avoidance policies. Many such restrictions occur because of limitations on the size (population) of geographic area which may be separately identified. In the case of microdata files, these restrictions, in addition to limiting the availability of data as such, sometimes make it impossible for the user to calculate sampling errors for the statistics of interest when such information is not provided by the releasing agency.

Recommendations

D 1. With respect to agency policies for releases, in statistical form, of information about individuals (natural persons), consideration should be given to the present apparent imbalance where there have been no instances of harm to individuals but several cases where requests for data have been denied. It is recommended that agencies review their policies to determine whether there are ways to respond more fully to user needs *without violating statutory requirements or risking harm to individual data subjects*. Some agencies may wish to try new data release procedures, such as controlled remote access to restricted microdata files, on a trial or experimental basis, with careful monitoring.

D 2. With respect to data for legal persons (corporations, etc.), both data subjects and data users have expressed some dissatisfaction with current agency disclosure-avoidance policies. The Subcommittee believes that continuing review of these policies is warranted, but it does not have any specific recommendations for change at this time.

E. Needs for Research and Development

Findings

1. Insufficient theoretical or empirical research has been carried out to determine the vulnerability of different classes of data to disclosure or the effects of disclosure-avoidance techniques on the utility of statistical data.

2. The Privacy Protection Study Commission (1977:587) has recommended, "That the National Academy of Sciences, in conjunction with the relevant Federal agencies and scientific and professional organizations, be asked to develop and promote the use of statistical and procedural techniques to protect the anonymity of an individual who is the subject of any information or record collected or maintained for à research or statistical purpose."

Recommendation

E 1. The Subcommittee would welcome a program of relevant research and development in the area of disclosure-avoidance techniques. Some particular areas that deserve attention are:

a. How disclosure risks in tabulations and microdata are related to varying sampling fractions.

b. How disclosure risks are related to the number of variables in the data base and to their individual and joint distributions.

c. Software systems for providing controlled online access to microdata files.

APPENDIX A

Statistical Disclosure-Avoidance Practices of Selected Federal Agencies

This appendix presents a description of the disclosure-avoidance practices of several Federal statistical agencies. The statements were prepared by the agencies and are presented here without comment. Agencies submitting statements are as follows:

STATEMENT BY THE BUREAU OF THE CENSUS

POLICIES AND PROCEDURES FOR AVOIDING DISCLOSURE IN THE RELEASE OF STATISTICAL TABULATIONS AND MICRODATA

A. Introduction

The Bureau of the Census operates under Title 13 of the U.S. Code, which prohibits the Bureau from making "any publication whereby the data furnished by any particular establishment or individual under this title can be identified." [1]

All data products are subjected to review to ensure conformance with established standards for the prevention of disclosure. Data released become available for purchase by anyone, which is also to say that data released to other Federal, State, or local governmental agencies must meet the same confidentiality standards imposed on data products prepared for general distribution.

Census Bureau disclosure rules may be discussed in terms of four different types of data sets: (1) tabulations from the censuses of population and housing, (2) tabulations from household surveys, (3) tabulations from economic censuses and surveys, and (4) public-use microdata.

B. Tabulations from Censuses of Population and Housing

The policy described here was that used for the 1970 census. Techniques to be used for the 1980 census are still under discussion.

1970 census summary data were primarily in the form of frequency tables with one or more dimensions, i.e., "count" data, using the terminology of Chapter III. The disclosure-avoidance techniques used in the 1970 census consisted of the suppression of the characteristics of small populations, i.e., populations of less than a certain threshhold size. This

[1] 13 USC 9. Section 8a of Title 13 does, however, permit individual information to be released to the person himself (or to his heirs). This service is primarily of use to persons who have no birth certificate or other legal proof of age or period of residence in the United States.

These sections of Title 13 do not apply to foreign trade statistics gathered under the provisions of USC 301 (13 USC 307).

approach was defined as "table suppression" in Chapter III. The rules were basically the same regardless of whether the data were disseminated in printed reports; on computer tape, on printouts, or any other form.

Data published from the 1970 census were based on either the "complete-count" part of the census—data obtained from every household—or the "sample" part of the census—data tabulated from the long-form questionnaires distributed to 20 percent of all households.

Suppression in complete-count data was based on a "rule of 5" for certain critical universes. The total population count in an area was never suppressed, but if there were fewer than 5 persons counted in an area then all distributions of characteristics about those persons were suppressed. Population characteristics cross-classified by race were subjected to an additional level of scrutiny: there had to be five or more persons in a racial category before data (e.g., an age distribution) could be shown separately for that race. For complete-count housing data, the rule of five was similarly applied to each race-of-head category, and also to distributions about homeowners, renters, vacant units, and a number of more specialized universes. A limited amount of complementary suppression was done to protect against disclosure by subtraction. For example, in a table where household size was shown for owners, renters, and the total, if there were fewer than five renters then data for owners were also suppressed to prevent derivation of renter data by subtraction from the total.

Data based on samples in the 1970 census were inflated to represent the total population. Thus, a person's response to a 20% sample question (e.g., education or income) was counted in tabulations as five responses on the average. Suppression thresholds were correspondingly inflated so that the rule of 5 became a rule of 25 for 20% sample data (representing five sample cases on the average) and 25 became the minimum number of persons or housing units in a critical universe for distributional statistics to be shown. Since there were actually two versions of the long-form questionnaire, one to 15% and the other to 5% of all households, there were also thresholds of 33 and 100 for those data, respectively.

1970 data were suppressed independently for different geographic areas. Thus the suppression of a figure for a small area was not allowed to preclude the publication of data for a larger area of which it was a part.

C. Tabulations From Household Surveys

Summary data based on small samples are not normally considered problematic with regard to disclosure. Sampling variability generally renders useless estimates based on a small number of cases, and consequently tabulations are not typically prepared for small areas or small populations. Published estimates from these surveys are nearly always rounded to the nearest hundred or nearest thousand.

In one survey where the sample size is large enough to support special tabulations for subcity areas, the rule of 5 actual cases, as applied in 1970 census sample data, is used. The inverse of the sampling rate is multiplied times five to derive a threshold which must be met by the total number of Blacks or persons of Spanish heritage before characteristics of those minorities are presented.

D. Tabulations From Economic Censuses and Surveys

Data generated about business firms in the economic censuses are generally in the form of magnitude data, as defined in Chapter III; for example, the total sales of all drug stores in a particular county. To avoid disclosure a cell suppression technique is used.

A dominance rule is employed in identifying sensitive cells: regardless of the number of respondents in the cell, if a small number (*n* or fewer) of these respondents contribute a large percentage (*k*% or greater) of the total cell value, then the cell is considered sensitive and is suppressed. The values of *n* and *k* are not revealed by the Bureau, since their publication would allow closer estimation of suppressed values, and would in turn require more suppression.

Cells found to be sensitive are necessarily suppressed, but so also are additional cells if their publication would allow the estimation of the sensitive value within certain bounds of equivocation. This may involve suppression of data for another industry within the same industry group, the industry group total itself, or a corresponding figure at another geographic level which may appear in a completely separate table. There may be more than one way to protect the sensitive cell. Data for large areas (e.g., a State) are given priority for publication over data for smaller areas (e.g., a county), and data for major industry groups are given priority over more specific industrial categories, where only one or the other may be provided. Further, an algorithm is employed to minimize the value of nonsensitive cells which must be suppressed to protect sensitive cells. These principles are further discussed in Appendix C.

While there may be a number of characteristics reported for a particular industry in an area, usually only one is designed as the key characteristic in determining dominance; if it is suppressed so also are the other characteristics. In the Census of Manufactures the key item is value of shipments; in the Census of Retail Trade, it is total sales. This dependence makes the disclosure analysis more manageable and avoids the possibility of inference from an unsuppressed characteristic to a suppressed characteristic.

Data are generally reported on an establishment basis, but disclosure analysis is performed on a company basis.

In a few cases where the data to be suppressed are of major significance the Bureau may ask a particular company to waive its right to confidential treatment and thus permit publication of the particular data.

Complete suppression is sometimes avoided by showing value ranges; for example, a table cell which would otherwise have been blanked out might carry a code indicating a range of $1.0 to $1.9 million dollars for value of shipments.

The count of establishments in an industry in any area is, by definition, not considered a disclosure. Further, the distribution of establishments by employment size class is not subjected to suppression. The size classes are sufficiently broad, however, that the upper limit of each interval is usually double or more the lower limit.

E. Public-Use Microdata

Microdata from censuses or surveys are released only to the extent that they cannot be identified to particular individuals. Identification is generally precluded by the absence of names and addresses, the release of records only for a small fraction of any population, and the exclusion of any information which would associate the respondent with a small geographic area. Microdata are not made available which could be matched against any known external files to identify the respondent. (For instance, the extent of data about identified business firms maintained by trade associations, regulatory agencies, and others has precluded the release of any microdata about businesses).

In general, public-use microdata files are designed to include all of the nongeographic information about the respondent captured in the census or survey. All characteristics have been recorded on the microdata in the same detail as encoded on the Bureau's computerized records, excepting only detail of high incomes (over $50,000 per year), Indian tribal identification, and detailed categories of group quarters. Any imputations for missing data are indicated as such to assist the user in evaluating the data.

The specific criterion regarding geographic identification is that no area with less than 250,000 population may be identified directly or indirectly. Thus, for example, State codes must be considered in conjunction with urban/rural codes and any other geographic identifiers on the file in determining what size of area is identified. Further, in microdata from a multistage sample survey, if the user can learn what areas were subject to sampling beyond a particular stage, then the 250,000 population criterion must be met in that part of the identified area that is known to have been sampled. The sequence of records within an identified area is scrambled to avoid any geographic inferences that could be made from the record sequence. Once one version of a file has been released, no other versions may be created with geographic detail which could be matched against the original file to violate the 250,000 criterion.

A total of six mutually exclusive public-use microdata samples, each including records for 1% of the population, were made available from the 1970 census. Each sample employed a different combination of subject matter and geographic options to meet as many types of needs as possible. A seventh special-purpose 1% sample was drawn covering only certain types of households. Requests for files which would have exhausted all of the available basic records of a particular type (20%, 15%, or 5%) were refused.

Microdata from each of the major household surveys conducted by the Bureau now generally become available for public use. These files are allowed to exhaust all sampled cases, but otherwise they observe the same geographic and other restrictions applied to census public-use samples.

There is no attempt to restrict the dissemination or types of use of these microdata files, and no files are released to Federal agencies or other special customers which are not also releasable to the general public.

STATEMENT BY THE BUREAU OF LABOR STATISTICS

LETTER FROM BARBARA BOYES, ASSISTANT COMMISSIONER, OFFICE OF SURVEY DESIGN TO JOHN A. MICHAEL, CHAIRPERSON, SUBCOMMITTEE ON DISCLOSURE-AVOIDANCE TECHNIQUES

This is in reply to your letter of May 20, in which you asked for a description of the disclosure-avoidance practices of the Bureau of Labor Statistics, for inclusion with the forthcoming report of the OMB Work Group on Disclosure Avoidance.

The Bureau's policy is one of informed consent: no identifiable data will be released without the express prior consent of the respondents.

There is no single nondisclosure rule for publication of BLS statistics. On the contrary, it is Bureau policy that publication criteria should take into account the special characteristics of each survey. Most surveys incorporate both nondisclosure rules and statistical reliability tests in their publication criteria.

There is one major category of BLS surveys which have disclosure rules that may be of interest to the Work Group. These are the major establishment surveys, which cover wages, employment and occupational injuries and illnesses. Most of them follow a threshold rule (at least three or four reporters per cell) and a dominance rule (one or two reporters may not account for more than 50% to 80% of the cell). Enclosure 1 is a list of those surveys, a brief description of each and the specific criteria applied to each.

The BLS has released two microdata tapes of the Consumer Expenditure Survey results, the "Diary Public Use Tape" and the "Summary Interview Tape." Because the Census Bureau conducts the survey under contract, the BLS is required by law to follow Census nondisclosure rules. Enclosure 2 gives the editing rules used on the tapes to avoid disclosure of individuals.

The BLS publishes a number of indexes, such as the Consumer Price Index, the Wholesale Price and Industrial Price Indexes, and the Employment Cost Index. These series all adhere to the "rule of three." In addition, the tests for sample adequacy that are applied at various levels of aggregation are such that disclosure problems are unlikely to arise.

A variety of other rules are applied to other types of surveys. For example, wage and benefit changes resulting from collective bargaining settlements are published either as a percentage change or as a cents-per-hour change, but not both, in order not to disclose the actual hourly rate for that specific group of workers.

BLS Nondisclosure Criteria, Major Establishment Surveys
(Enclosure 1)

The *Employment and Wages* series is derived from the file of all establishments covered under the State Unemployment Insurance programs. Tabulations give the number of employees and total wages for each State by industry and size of reporting unit. The number of reporting units is also shown but in much less detail, i.e., State by major industry, or State by size class.

Threshold: 3 reporting units (firms or establishments)

Dominance: 2 firms at 80%

The *Industry Employment Statistics* series is derived from a large sample of establishments and consists of tabulations of number of employees, average earnings and hours for detailed industries on the national level. Average earnings and hours are also shown for States and areas with less industry detail

Threshold: 3 firms

Dominance: 2 firms at 80%

The *Occupational Employment Statistics* series is published by individual States from a sample survey of establishments reporting on the occupational structure of selected industries.

Threshold: 3 firms

Dominance: 1 firm at 50% or 2 at 75%

The *Industry and Area Wage* surveys supply averages and distributions of wage and salary rates for selected occupations or industries. Data usually refer to specific SMSA's or ad hoc aggregations of areas Other detail may appear in the publication, e.g manufacturing/non-manufacturing or part-time/full-time categories.

Threshold: 4 establishments or 7 (weighted) workers

Dominance: 1 establishment at 60%

The *Occupational Safety and Health Survey* pro duces national injury and illness rates by industry and employment size of establishment.

Threshold: 3 firms

Dominance: 1 firm at 50% or 2 at 75%

Confidentiality and Tape Content

(Enclosure 2: Memorandum dated September 22, 1976 from Eva E. Jacobs, Chief, Division of Living Conditions Studies to John Layng, Assistant Commissioner for Prices and Living Conditions)

In determining the characteristics and the form of the characteristics on the public use tapes we have been guided by the principle of providing as much detail as possible within the limits of protecting the confidentiality of the data. However the requirements for confidentiality are not specific except for (1) Census requirements that areas under 250,000 population not be identified and (2) BLS' Commissioner's guidelines which forbid the identification of individual data.

In general there will be a tug of war between users who want every bit of information and the agency which is committed to preventing disclosure. The standards we have adopted are pragmatic and have resulted from examining counts of respondents with what might be identifying characteristics when combined with the amount of geography being shown.

The following characteristics were limited:

1. *Geography.*—No individual areas are identified. Region, size of area, inside, outside SMSA are shown.

2. *Income.*

   a. Actual income except for under $2,000 and $35,000 and over.

   b. Sources of income.

Earnings of head, spouse and other. Other income is aggregated into 4 groups.

(1). Social security and railroad retirement,

(2). Government retirement and

(3). Interest, dividends, rent and royalties.

(4). Public assistance and other.

For the diary, family income was collected on an aggregated basis. The interview collected income in detail. For the summary tape, the income will be shown the same as the diary. For the later detail tape, a decision will have to be made as to whether the individual items such as public assistance, interest, dividends, unemployment compensation, should be shown. The number of respondents with public assistance outside SMSA's in rural areas, for instance, is very small.

3. *Family size.*—Actual up to 6, then 7+.
4. *Age.*—Actual up to 74, then 75+.
5. *Race.*—"Other" has been combined with white because there are a very small number. This leaves the "black" category as a separate group.
6. *Marital status*—Married, other. We are not showing widowed, divorced, never married.

Does this approach meet with your approval?

STATEMENT BY INTERNAL REVENUE SERVICE

Disclosure Policies and Procedures with Respect to Statistical Information

The Internal Revenue Service prepares and releases in its annual Statistics of Income publications aggregated data derived from samples of income tax returns of individuals, corporations and unincorporated businesses, as required by section 6108 of the Internal Revenue Code (26 U.S.C. 6108 as amended by the Tax Reform Act of 1976). On a less frequent basis, the Service also prepares and publishes similarly derived statistics for fiduciaries, estates, gifts, and domestic corporations with foreign operations.

Section 6108 of the Code further provides in subsection (c): "No publication or other disclosure statistics or other information . . . shall in any manner permit the statistics, study, or any information so published, furnished, or otherwise disclosed to be associated with, or otherwise identify directly or indirectly, a particular taxpayer." In implementing this provision of the Code with respect to statistical tabulations (aggregated data) the Service follows a rule of 3 with respect to data on a National or State level, such that data based on fewer than three returns are suppressed before they are released. In the case of tabulations with geographic detail below the State level, a rule of ten is followed with data based on less than ten returns suppressed.

Subsection (b) of section 6108 provides that special statistical studies may be prepared and furnished to requesters on a reimbursable basis. On the basis of this provision, the Service can provide requesters, for a fee, special statistical tabulations and in addition a computer tape file containing unaggregated or microdata data without information that would identify specific taxpayers. This file is the National Individual Tax Model, which consists of a subsample of the regular Statistics of Income sample of individual income tax returns. The identifying information deleted from the file consists of Social Security Number (the numbers of both husband and wife in the case of joint returns) and geographic codes identifying State or Internal Revenue District.

One other microdata set—the State Individual Tax Model—is made available to requesters. This set is partitioned, based on the taxpayer's address, into subfiles for each one of the States and the District of Columbia. To maintain reliability of estimates, each of the State subfiles is based on the full Statistics of Income sample rather than a subsample. In releasing a subfile for any particular State, Social Security Numbers are deleted and, in addition, return records with high incomes ($200,000 or more) are deleted completely to preclude the possibility that such returns, particularly those with very high incomes (which are selected for the sample at a 100 percent rate), could be associated with well-known individuals residing in a particular State.

STATEMENT BY NATIONAL CENTER FOR EDUCATION STATISTICS

Disclosure-Avoidance Practices

The National Center for Education Statistics (NCES) has the responsibility to "collect, collate, and from time to time, report full and complete statistics on the conditions of education in the United States; conduct and publish reports on specialized analyses of the meaning and significance of such statistics" (Statute 501 of P.L. 93–380). NCES also has responsibility to protect the confidentiality of certain information pertaining to individuals and institutions. While each set of data is regarded as unique, thus requiring its own, separate treatment, disclosure-avoidance practices in NCES can be conveniently grouped for purposes of this report as follows:

Deletion of Identifiers
Cell sizes
Collapsing of Report Data
Professional Review

Unless otherwise noted the disclosure-avoidance practices described below apply to both statistical tabulations and microdata tapes (computerized records of individual statistical units).

Deletion of Identifiers and traceable data (e.g. geographic location) is an NCES practice in dealing with data which might be used separately or in association with still other data to indicate information about persons (individual or organizational) regarded as confidential.

Cell size in some instances, has relevance to disclosure-avoidance practices. The "rule of three" (involving fewer than four cases) involves the deletion of confidential information about three or fewer persons before tabulations and microdata files are released.

Collapsing of Report Data occurs in some NCES statistical reports by combining cells, lines or columns of information, into larger class intervals or broader groupings of characteristics.

Professional Review by staff responsible for the data is required of all NCES data releases to discern and note possible disclosures of confidential information not detected through other safeguards.

STATEMENT BY NATIONAL CENTER FOR HEALTH STATISTICS

POLICIES AND PROCEDURES TO AVOID INADVERTANT DISCLOSURES THROUGH PRINTED PUBLICATIONS AND PUBLIC-USE MICRODATA TAPES

A. Introduction

The National Center for Health Statistics is authorized under the Public Health Service Act (42 USC 242k, Sec. 306(b)(1)) to collect statistics on the extent and nature of illness and disability of the population of the United States; the impact of illness and disability on that population; environmental, social, and other health hazards; determinants of health; health resources; utilization of health care; health care costs and financing; family formation, growth, and dissolution; and births, deaths, marriages, and divorces. Such data are obtained through a variety of means—through State vital statistics registration, from large-scale population surveys, surveys of institutions and practitioners, State licensing programs for practitioners and institutions, encounter forms and abstracts from health care practitioners, reports from agencies, and compilations of other national organizations.

As is noted in Chapter II, the Public Health Service Act also requires that the confidentiality of information obtained by the Center be protected: Data may be used only for the purpose for which it was collected, and data identified with an individual or establishment may be disclosed only with the consent of that person or establishment or the provider of the data. (Section 308(d)) Departmental regulations have not yet been promulgated to implement this Section, but its meaning is still quite clear in the absence of regulations.

NCHS is in process of reviewing, revising, and strengthening its internal regulations regarding the avoidance of inadvertent disclosures of confidential information. But until such new regulations are published, those contained now in the Center Manual continue in effect.[1]

B. Release of tabulations

The Manual issuance section speaks to the concern over the "publication" of statistical data that unintentionally identifies persons or establishments, or displays measures which a reader can ascribe to individual persons or establishments. The following rules, with modifications. are set forth:

"*Rule of Three.*—Except as specially otherwise provided, published tables should show no data in cells for which the reporting units are less than 3 in number. Care must be exercised that data do not appear in "residual" cells, or can be derived for such cells by subtraction, if the residual represents less than 3 reporting units.

"*Modifications to Rule of Three.*—1. In some cases it is feasible to present separate data for two or even one respondent. One group of such cases includes presentation of *rate* data, when there are no collateral data to further identify the individual reporting unit. For example, assuming the absence of other identifying information, it would be acceptable to show within a single 2-way cell these data for 3 hospitals not otherwise described:

Average length of stay
7.2
7.8
7.9

"The guideline here is that even though data for each of three single hospitals are shown, this publication does not identify the individual hospital.

"2. Tables may show simple *counts* of number of persons, even though the number in a cell is only 1 or 2, provided the classifying data are not judged to be sensitive in the context of the table. For example, publication of counts of health manpower personnel by occupation by area are considered acceptable, if not accompanied by other distinguishing characteristics, or other cross-classifications which have the effect of adding descriptive information about the same persons. But publication of counts of personnel for a specified occupation by area by income is not acceptable for cells of less than 3 persons, because that would reveal sensitive income data.

"3. In some situations, it may not be acceptable to publish a cell which contains data for as many as 3 or even more reporting units. For example, suppose there are 5 recognizable establishments which con-

[1] NCHS Staff Manual Guide. General Administration No. 3, Supplement No. 3, June 24, 1974.

stitute the membership of a cell, but one of the 5 represents 90 percent of the activity in that cell. It would be undesirable, and possibly illegal, to publish for that cell the proportion of discharges which were not alive, since that would permit a highly accurate estimate of the rate for the individual establishment. A guideline for this situation is to suppress if one establishment accounts for as much as 60 percent of the magnitude for the cell.

"*Impact of External Data.*—It is clear that knowledge of several descriptive attributes of a given person generally makes it more likely that the person can be identified than if only a single descriptor is known. Furthermore, since there are many files, both governmental and non-governmental, containing information about persons, there is always the possibility that cross-tabulation of data from one file with those from another file might yield sufficiently unique categories that one or more persons could be identified from the merged files, even though neither alone would permit that. (Indeed, some students argue that given enough money and time, it is always possible to break any camouflage of identification.) NCHS guidelines for presentation of statistical data require only that the NCHS data themselves do not reveal identity. It is not necessary to consider whether merges of real or hypothetical external files would compromise security of the information; *except* that NCHS will be alert not to publish cells for which there is *common knowledge* of other characteristics which would permit matching of data for individuals. For example: NCHS should not publish or release information previously considered confidential, for a cell which was described as relating to (1) a male, killed by gunfire, in Dallas, Texas on November 22, 1963; nor (2) average salary of nurses in privately-owned hospitals with 1,000 or more beds in any specified community.

"A special situation prevails in the vital statistics area, where the State is the collector under its own law. NCHS uses the data under a contractual arrangement with the State, which fills the role of respondent in this context. NCHS does abide by the terms of the contracts, although it can exercise no control over how the State manages other confidentiality aspects of the vital records. Under the terms of the contract, NCHS will not permit access to individual records nor will it give the "key" (certificate number) to individual records to anyone without the expressed written consent of the State (registration area). Nevertheless, it has been a long-standing tradition in the field of vital statistics not to suppress small frequency cells in the tabulation and presentation of data. For example, it has been considered important to know that there were two deaths from rabies in Rio Arriba County, New Mexico in a given year, or that there were only one infant death and two fetal deaths in Aitkin County, Minnesota. These types of exceptions to general NCHS practices in other programs are followed because they have been accepted traditionally, and because they rarely, if ever, reveal any information about individuals that is not known socially.

"*Rule of Reason at Editing Stage.*—It is not expected that NCHS rules for release of data be so repressive as to attempt to remove all possibility of identification of individual reporting units, or of revelation of restricted information concerning an individual reporting unit, should a probing investigator choose to expend unlimited resources to secure such information. It is expected in addition to adherence to the guidelines stated herein that Division and Office reviewers of NCHS reports be ever conscious of the Center's commitment to protect respondents, and take any special *ad hoc* action which appears necessary, and similarly, that final editors be alert to call attention to situations that appear questionable."

C. Release of Microdata

NCHS has a rapidly growing program of providing data from its activities to researchers on magnetic tapes, some having summary data, and some with microdata. A publication states the Center's policies on release of microdate.[2] Its gist is summarized in the following policy statement:

"Within prevailing ethical, legal, technical, technological, and economic restrictions, it is the policy of the National Center for Health Statistics to supplement its central programs of collection, analysis and publication of statistical information, with procedures for making available, at cost, transcripts of data for individual elementary units—persons or establishments—in such form as will not in any way compromise confidentiality guaranteed the respondents."

The public use data tapes produced by the Center are catalogued in a Center publication which is updated annually.[3] In keeping with the law's require-

[2] *NCHS Policy Statement on Release of Data for Individual Elementary Units and Related Matters.* DHEW Publication No. (HSM) 73–1212. 1973. USGPO, Washington, D.C.

[3] *NCHS Standardized Micro-Data Tape Transcripts.* DHEW Publication No. (HRA) 76–1213. 1976. USGPO, Washington, D.C.

ment that the data be used only for the purposes for which they were collected, purchasers of micro-data tapes are required to complete the following, which is part of the Order Form:

"Individual identifiers have been removed from the micro-data tapes available from NCHS. Nevertheless, under section 308(d) of the Public Health Service Act, such information may not be used for any purpose other than the purpose for which it was supplied. The information on the micro-data tapes available for purchase was supplied to NCHS for statistical research and reporting purposes. It is necessary therefore that the individual ordering such micro-data tapes sign the following assurance:

> "The undersigned gives assurance to NCHS that individual elementary unit data on the micro-data tapes being ordered will be used solely for statistical research or reporting purposes.
>
> Signed: ______________________
> Title: ______________________
> Organization: ______________________
> Date: ______________________"

The Manual issuance cited above also sets forth the following:

"*Micro Data Tapes.*—On all magnetic tapes of micro data which are released outside the NCHS, geographic identification must be deleted for all areas below the *State* level which contain fewer than 250,000 inhabitants in the most recent official population Census. The most likely procedure for accomplishing this is to substitute for all smaller areas a new code, "Rest of State". Codes for such characteristics as population density or SMSA, non-SMSA, but which do not identify individual areas, may appear on the tapes for areas with less than 250,000 inhabitants.

"It may be necessary to suppress certain other classifying codes in special situations, or in establishment data, although in general the geographic suppression indicated above will be considered a sufficient protection for person or household data."

STATEMENT BY SOCIAL SECURITY ADMINISTRATION

Policies and Procedures for Avoiding Disclosure in the Release of Statistical Tabulations and Microdata

A. Introduction

There are several sources and uses for data released by the Social Security Administration (SSA). Some SSA publications include statistical tabulations containing program data either obtained directly from records used to administer social security and other programs or compiled from samples of these administrative records. Some publications are based on surveys in which data are collected directly from actual or potential participants in social security programs. In addition, SSA makes microdata files, i.e., tape files of individual records without identifiers, based on program and/or survey data, available to outside researchers.

Legal requirements for confidentiality in such tabulations and microdata releases are based on Regulation Number 1, promulgated under Section 1106 of the Social Security Act, and on general statutes such as the Privacy and Freedom of Information Acts. As a matter of policy, the Social Security Administration has consistently taken a strong position on the confidentiality of information about individuals participating in social security programs.

To comply with these legal and policy standards in the release of tabulations and microdata for statistical purposes, SSA has taken a "two-tier" approach. In cases where disclosure risks are considered to be minimal or non-existent, tabulations and microdata files are released without restrictions on their use. In other cases, where public policy requirements are considered to outweigh small but non-negligible disclosure risks, releases are made only on a restricted basis, under written agreements covering the use and safeguards of the material released. Specific examples are presented below.

B. Release of Tabulations

A comprehensive set of guidelines for preventing disclosure in tabulations containing program data has been developed to control disclosure in unrestricted releases based upon 100 percent data (see attachment). On the other hand, when tabulations derived from complete program data are supplied exclusively to the Bureau of Economic Analysis (BEA) for its internal use to develop regional income estimates, the figures delivered by SSA are not modified according to these guidelines. However, BEA is responsible both for internal security and for the release of its results in a way that will not identify specific individuals. Source documents are returned to SSA after BEA has extracted the information it needs.

There are generally no restrictions placed upon the release of tabulations based upon sample data with limited geographic information (for example, national and regional only). Because of the uncertainty about whether or not a particular member of a cell is included in a sample, especially when the sampling fraction is small, fewer restrictions are necessary for sample data than for 100 percent data in the release of figures corresponding to the same cell. Even though detailed geographical information may be present, for example, there are no restrictions on tabulations based on the 1-percent file from the Continuous Work History Sample (CWHS). In particular, earnings information at the State and metropolitan area levels was published without suppression or disturbance in *Earnings Distributions in the United. States, 1969*.

On the other hand, in summary tabulations prepared by BEA from SSA's 10-percent CWHS file, some restrictions are applied. Data in all tables are rounded to the nearest 100 workers, and tables are printed only when the total number of workers in the sample is 400 or more. Data on industry by county are suppressed when such cells are dominated by a small number of establishments.

C. Release of Microdata

When microdata based on small samples with limited geographic information are to be released, the files are first reviewed to suppress unusual values or combinations of values, or to present certain items in class interval rather than exact form. The records are then released to users without restrictions. Examples of such microdata files available for public use are those derived from the Longitudinal Retirement History Survey, the Survey of Low Income Aged and Disabled, Disability Surveys, and a 1973 CPS-IRS-SSA Match Study.

The CWHS microdata files, which contain more geographic and other detail for individuals, are released only subject to restrictions covered by written agreements. Files from the 10-percent samnle have been released only to the Census Bureau and BEA. Starting in 1976, files from the 1-percent and 0.1-percent samples have been released only subject to execution of a "conditions of use" agreement in which the recipient agrees, among other things:

- To use the files only for statistical and research purposes specified in the agreement.
- To refrain from trying to identify, for any purpose, specific individuals or employers.
- Not to release the files to any other organization or individual unless authorized by SSA.
- Not to publish or otherwise release tabulations or listings which might reveal information about identifiable individuals or employers.

In addition, the following precautions are taken:

- Files are tailored to individual user requirements, i.e., only the specific data items needed by the user are included in the file released to him.
- Random noise is introduced into the earnings information.

Some of the data in CWHS files for 1976 and subsequent years are considered to be tax return information, as defined in the Tax Reform Act of 1976, and are therefore subject to the disclosure provisions of the Internal Revenue Code, as amended by that Act. Therefore, policies for release of these files are undergoing further review.

One 100-percent microdata file was released by SSA for research and statistical use. This was an extract from our Chronic Renal Disease file that was released to an HEW contractor. Specific dates of events, beneficiary and provider ID's and other information likely to disclose individual identities were removed from the records. Conditions similar to those described above for CWHS releases were agreed to by the recipient.

Guidelines for Preventing Disclosure in Tabulations of Program Data

(Attachment to SSA Statement. Memorandum dated February 16, 1977, from John J. Carroll, Assistant Commissioner of the Office of Research and Statistics, to ORS Executive Staff.)

The following guidelines are primarily of concern to the Divisions of Health Insurance Studies, OASDI Statistics and Supplemental Security Studies. They apply to the release to non-SSA users, in published or unpublished form, of statistical tabulations of SSA program data based on complete counts for individuals or for groups of beneficiaries within a family.

Good statistical practice, as well as provisions of statutes and regulations, require that we strike a careful balance between the protection of individual privacy and the needs of users for data about social security programs. These guidelines have been developed on that basis. Directors of the divisions mentioned above are requested to distribute copies of these guidelines to staff members responsible for the release of program data, and to instruct them to follow the procedures in the guidelines.

There may be some areas in which immediate full compliance would be difficult, and there will undoubtedly be some situations not specifically covered by the guidelines. All questions of this nature should be referred to the Chief Mathematical Statistician. I am also requesting that he review publications and other releases from time to time to assure that suitable disclosure prevention procedures are being used.

A. Introduction

SSA Regulation No. 1 permits the release of "statistical data or other similar information not relating to any particular person which may be compiled from records regularly maintained by the Department." Under this authority the Office of Research and Statistics releases a variety of tabulations, in both published and unpublished form, to users outside of SSA.

The phrase "not relating to any particular person" is taken to mean that SSA should not release any tabulation that makes it possible for a user to identify a particular person included in the tabulation and thereby to obtain additional information about that person. Such inadvertent release of information about individuals is called "disclosure."

Strictly speaking, there is no such thing as absolute protection against disclosure in statistical tabulations. Any tabulation provides some information about persons known to be included in it. What we must provide, then, is a *reasonable* degree of protection against the disclosure of precise information about any individual, especially when such disclosure is potentially embarrassing to that individual.

ORS divisions have used several different rules and procedures to avoid disclosure. The guidelines that follow were developed in response to an expressed need for uniform standards or principles concerning the kinds of disclosures that should be avoided and the appropriate methods of preventing such disclosures.

B. Scope

These guidelines apply only to the release to non-SSA users, in published or unpublished form, of statistical tabulations of SSA program data based on complete counts for individuals or for groups of beneficiaries within a family. The release of microdata files and of sample tabulations is not covered.

Separate standards are provided for *count data*, i.e., numbers of persons or other beneficiary units, classified by characteristics such as age, sex, race and residence; and for dollar amounts, i.e., total or average benefits for various classes of beneficiaries.

For each of these two categories, *basic rules* are provided describing the kinds of disclosure that must always be avoided. Staff preparing tabulations are also encouraged to take steps to avoid less obvious possible disclosures, especially when dealing with more sensitive classes of information.

A special rule is provided for those instances in which an outside user requests SSA to merge individual earnings and/or benefit data with information provided by him for specific individuals, and to release tabulations based on the merged records.

A brief discussion of different methods of preventing disclosure is included. No single method is recommended in preference to all others. The choice will depend on the structure of the tables and on the nature of the data processing systems being used to produce them.

C. Count Data

1. *Basic rules.—*
   a. No tabulation should be released showing distributions by age, earnings or benefits in which the individuals (or beneficiary units, where applicable) in any group can be identified to
      (1) an age interval of 5 years or less.
      (2) an earnings interval of less than $1,000.
      (3) a benefit interval of less than $50.
   b. For distributions by variables other than age, earnings and benefits, no tabulation should be released in which a group total is equal to one of its detail cells. Some exceptions to this rule may be made, on a case-by-case basis, when the detail cell in question includes individuals in more than one broad category.

The rationale for these rules is that if a user can identify an individual as being a member of the group for which the distribution is shown, the fact that that individual is also known to be in the detail cell or combination of adjacent cells will provide the user with additional information about him.

2. *Examples for basic rules.—*

Rule a.

Number of beneficiaries by monthly benefit amount, by county

	Monthly benefit amount						
County	$0–19	$20–39	$40–59	$60–79	$80–99	$100+	Total
A ----	2	4	18	20	7	1	52
B ----	—	—	7	9	—	—	16
C ----	—	6	30	15	4	—	55
D ----	—	—	2	—	—	—	2

The distributions can be shown for counties A and C, but not for B and D. For county D, there is only one non-empty cell, and a beneficiary in this county is known to be receiving benefits between $40 and $59 per month. For county B, there are 2 non-empty cells, but the range of possible benefits is less than $50, i.e., from $40 to $79 per month.

Rule b.

Number of beneficiaries by race, by county

	Race			
County	White	Black	Other	Total
A ----	15	3	—	18
B ----	30	—	—	30
C ----	72	20	2	94
D ----	27	—	2	29

The distributions can be shown for counties A, (and D, but not for B. In county B, the number c white beneficiaries is equal to the total.

3. *Additional restrictions.—*Except as noted fo age, earnings and benefit distributions, the basic rul does not prohibit empty cells as long as there ar 2 or more non-empty cells corresponding to a mar ginal total, nor does it prohibit detail cells with onl one person. However, additional restrictions (se below) should be applied whenever the detailed clas sifications are based on sensitive information. Th same restrictions should also be applied to non sensitive information if it can be readily done an does not place serious limitations on the uses of th tabulations.

Sensitive information includes, but is not necessarily limited to, the following:

- Race
- Diagnosis of medical condition
- Program entitlement, as follows:
  Title II—disability
  Title XVI—all categories
  Title XVIII—disability

Additional restrictions may include one or more c the following:

(a) No empty cells. An empty cell tells the use that an individual included in the margina total is *not* in the class represented by th empty cell.

(b) No cells with one person. An individual in cluded in a one-person cell will know that n one else included in the marginal is a membe of that cell.

(c) No tables for which any of the restrictions i the basic rule and items (a) and (b) directl above would be violated by tables directl derivable (usually by subtraction) from th tables released.

D. Dollar amounts

1. *Basic rule.—*An individual's (or couple's) exac benefits should never be disclosed. Disclosure ca happen in two ways:

(a) Release of an average or total amount for publication cell with only one member. (Re vealing average or total benefits to the neares whole dollar for a one-person cell will be con sidered the same as revealing exact benefits

(b) Release of an average or total amount for publication cell if the individual benefit amour has known upper and/or lower limits, and a members are at one of those limits.

Example: The maximum benefit for a certain program is $230 per month. If the average benefit for a particular cell is $230 per month, then it will be known that anyone included in that cell is receiving that amount.

2. *Additional restrictions.*—Further restrictions should be applied under the same general conditions as those described for count data. Additional restrictions may include:

(a) No publication of total or average amounts for cells containing only two members.

(b) No publication of total or average amounts for cells if the information provided, in conjunction with known upper and lower limits, would make it possible to deduce that all persons in that cell were receiving benefits within a restricted range, e.g., a range of less than $50. Examples of such disclosures and the procedure for determining when they occur are shown in Exhibit 1.

E. Special Rule for SSA Data Merged with User Data

Special care is necessary to avoid disclosure when tabulations for release are based on SSA program information such as earnings and benefit data merged with individuals' records containing other data supplied by researchers outside of SSA. This is because we know that the outside user who supplied the individual records to SSA has access to considerable information about each individual included in the tabulations and therefore can readily identify individuals in small tabulation cells.

1. *Basic rules.*—In tabulations based on merged SSA and user data, no SSA data may be provided for groups of fewer than five persons formed on the basis of information provided by the user. For groups of five or more persons, SSA data may be presented subject to the restrictions described in the previous sections for counts and dollar amounts.

2. *Exception.*—Disclosure of the fact, date and circumstances (generally interpreted to mean location) of death of an individual is permitted by SSA Regulation 1. Therefore, no restrictions on tabulations are required if the only effect would be to disclose this kind of information.

F. Methods of Preventing Disclosure

As stated earlier, no single method of preventing disclosures is recommended. The choice will depend largely on the techniques used to produce the tables and on the frequency with which disclosure situations are expected to occur.

Methods of preventing disclosure fall into two broad categories:

- Suppression and grouping of data
- Introduction of error

Each of these is discussed further below.

1. *Suppression and grouping of data.*—Suppression consists of simply not showing the values for certain cells of a table. Usually the numerical values (including zeroes) that are suppressed are replaced by a symbol footnoted to explain that the item was suppressed to avoid disclosure.

Grouping consists of combining cells (or lines or columns or other units) of a table to produce a revised table without disclosures.

The main difficulty with suppression and grouping techniques is that they must be applied with great care to avoid "complementary disclosure," i.e., a situation where the elements of the table suppressed or grouped can be derived from the information remaining in the table. As a simple illustration, consider a table containing a line of data for each county in a State, and a line with the corresponding State totals. If the data for a single county are suppressed to avoid disclosure, the user can derive them by adding the data for the remaining counties and subtracting from the State totals. To avoid such complementary disclosure, it would be necessary either to suppress data for two or more counties in the State, or to group data for two or more counties.

If disclosure problems are frequent in a particular set of tables, the job of making the necessary groupings and/or suppressions may be very laborious. Furthermore, the use of these procedures on an ad hoc basis does not lend itself readily to automation.

Instead of attacking the specific disclosure problems described in these guidelines each time they occur, it has been helpful in some cases to begin by applying more general rules which eliminate most of the disclosures. For example, in a table presenting selected data on benefits by State and county we might observe the following rules:

a. Do not present data for any individual county with fewer than 50 (or some other number) beneficiaries.

b. Do not suppress data for a single county in a State. If suppression is used, data for two or more counties must be suppressed. Alternatively, small counties may be grouped so that no data are shown for counties or groups of counties with fewer than 50 beneficiaries.

If the minimum number of beneficiaries has been appropriately chosen, application of these general rules will eliminate most potential disclosures; the few remaining can be dealt with easily.

2. *Introduction of error.*—The probability of disclosure can be reduced by introducing error or "noise" into the data. The error may be introduced into the records for individuals prior to tabulation, or it may be introduced into the cells of the tabulations. The error may be introduced in a purely systematic way, as in ordinary rounding, or it may contain some element of randomness.

Many different methods of introducing random error have been used in practice. As one illustration, consider the following method of rounding all cells of a table so that they end in 0 or 5. Each detail cell value not ending in 0 or 5 is rounded to the next higher or lower number ending in 0 or 5, as follows:

Ending digit	Probability of	
	Rounding down	Rounding up
1	4/5	1/5
2	3/5	2/5
3	2/5	3/5
4	1/5	4/5

The actual direction of rounding for each cell is determined by the appropriate use of random numbers. This technique eliminates the need for grouping and suppression of count data in tables, as a 0-cell in the resulting contaminated table may or may not represent a 0-cell in the original table. It is important, of course, that users be informed that random errors have been introduced.

There are many variations and refinements of the technique illustrated. A "controlled" random procedure may be used to minimize the distortion of totals and subtotals derived from the detail cells. Rounding does not have to be to numbers ending in 0 and 5; it may be sufficient to round all cells to even numbers. The errors introduced may be either additive (as in the rounding process) or multiplicative. In either case, the expected value of any cell should be its original value.

These techniques can be automated. The initial investments of programming effort may be substantial, but once the system is developed, little if any further attention to the disclosure problem is needed. The obvious disadvantage of introducing errors is that the user must deal with data that are less precise.

The Division of Supplemental Security Studies has developed and successfully tested a program for random rounding of individual tabulation cells in their semi-annual tabulations of Supplemental Security Income State and county data.

G. Bibliography

Further discussion and illustration of techniques for identifying and preventing disclosure may be found in the bibliography attached. Copies of items listed may be obtained by calling 673-5727.

1. Barabba, Vincent R. and Kaplan, David, L., "U.S. Census Bureau Statistical Techniques to Prevent Disclosure—The Right to Privacy vs. the Need to Know," paper presented at the 40th Session of the International Statistical Institute, Warsaw, Poland, September 1975.

2. Dalenius, Tore, "The Invasion of Privacy Problem and Statistics Production—an Overview," *Statistisk tidskrift* 1974; 3.

3. Fellegi, I. P., "Controlled Random Rounding," *Survey Methodology,* 1975, Vol. 1, No. 2.

4. Fellegi, I. P., "On the Question of Statistical Confidentiality," *Journal of the American Statistical Association,* 1972, pp. 7–18.

5. National Central Bureau of Statistics, *Confidentiality in Statitical Tables,* Stockholm, Sweden, 1974. See especially Chapter 4, "Methods of Solving Confidentiality Problems in the Production of Tables."

6. Newman, Dennis, "Techniques for Ensuring the Confidentiality of Census Information in Great Britain," paper presented at the 40th Session of the International Statistical Institute, Warsaw, Poland, September 1975.

Exhibit 1
Disclosure of Benefit Ranges

A. Introduction

When the upper and lower limits of possible benefit payments to individuals are known, publication of the total or average benefits for a particular group can sometimes reveal that the benefits for all members of that group lie within a range of values that is narrower than the range between the known upper and lower limits. Release of information under these circumstances is a form of disclosure, even though the exact amount may not be revealed for any individual.

This note tells how to detect the existence of such disclosures.

B. Notation

For a particular group of individuals receiving benefits under some program, we assume that the following data are being considered for release:

N = number of individuals receiving benefits

A = average benefit amount

and that the following values are generally known:

U = maximum possible payment

L = minimum possible payment (it is assumed that $L > 0$)

$R_0 = U - L$

The following data are only available internally:

X_U = the largest payment to any member of the group

X_L = the smallest payment to any member of the group

$R = X_U - X_L$

C. External disclosure

External disclosure occurs when someone not a member of the group can determine from the data released that the largest possible (not the actual) range of benefits for that group is smaller than R_0.

Disclosure will *not* occur whenever

$$\frac{U + L(N-1)}{N} \leq A \leq \frac{L + (N-1)U}{N}$$

If $A \leq \dfrac{U + L(N-1)}{N}$

Then it will be known that

$$X_U \leq AN - L(N-1) < U$$

If $A \geq \dfrac{L + (N-1)U}{N}$

Then it will be known that

$$X_L \geq AN - U(N-1) > L$$

Example:

Suppose $N = 10$, $L = \$10$, $U = \$90$

No disclosure will occur if $\$18 \leq A \leq \$82$

If $A = \$12$, we know that

$$X_U \leq \$30$$

and, of course, if $A = \$10$, we know that

$$X_U = \$10$$

If $A = \$85$, we know that

$$X_L \geq \$40$$

D. Internal disclosure

A person who is a member of the group, in addition to knowing N, A, L, and U, will know the value of his own payment, X. He will be able to calculate the average payment, A′, for the remaining members of the group, i.e.

$$A' = \frac{NA - X}{N-1}$$

No disclosure to that individual will occur as long as

$$\frac{U + L(N-2)}{N-1} \leq A' \leq \frac{L + (N-2)U}{N-1}$$

To determine whether or not internal disclosure can occur for a particular group, it will be sufficient to make this test for $X = X_U$ and $X = X_L$.

STATEMENT BY STATISTICAL REPORTING SERVICE, USDA

A. Unintentional disclosure through published tables.

The Statistical Reporting Service (SRS) of the U.S. Department of Agriculture (USDA) collects, processes, and disseminates agricultural statistics. Nearly all data are collected from respondents under a voluntary system, without statutory reporting requirements. Therefore, it is imperative that each respondent's confidentiality be maintained.

When data are aggregated at the county level it is possible that an individual operation could be disclosed. Aggregated data at the State level can also reveal an individual operation for a specialized agricultural commodity.

Each State Statistical Office (SSO) is responsible for the detection of any disclosure problem which might occur in the data for that State. If a disclosure problem is detected, the SSO will recommend that the State's estimate for the commodity in question not be published separately but be combined into an "other States" category.

Data submitted for publication in a national release are also checked by the commodity statistician responsible for the U.S. estimate. The States and the national commodity statisticians follow the "rule of three" (do not publish separate State estimate if less than 3 operations are included in the total) and the "sixty percent rule" (do not publish separate State estimate if one operation has 60% or more of the total). State estimates withheld from separate publication under these rules are combined into an aggregate labeled "other States" and appropriately footnoted. If only one State total is withheld under these rules, the commodity statistician will select another State and combine the totals for the two, thus withholding individual totals for two States in order to prevent disclosure.

The only exceptions permitted to the policy forbidding the publication of potentially identifiable data are with the written permission of the individual operation(s) concerned.

B. Unintentional disclosure through release of microdata tapes.

SRS does not release identifiable data on individual operations to other agencies. Aggregated county estimates data are released, and the same rules apply as for the published tables. SSO's are responsible for determining when data should be merged to protect confidentiality.

APPENDIX B

Protecting Data in Computer Systems

Mervyn R. Stuckey, Statistical Reporting Service, USDA

Introduction

The computer revolution has changed data collection and analysis. Due to the speed of the computer, data are often available for analysis in a relatively short period of time after collection. Files can be linked, and much can be learned about an individual. Vast amounts of data can be stored and made readily available. Telecommunications allows us to link computers and thereby multiplies the capability to exchange and share data.

The recognition of data as a resource has also created concern among management. Losses of data can be detrimental to the organization holding the data, as well as to the individual or group included in the data. Data resource management is not a new field of endeavor, but it is receiving more attention now than ever before. The Privacy Act of 1974 brought the data confidentiality issue to the forefront. The National Bureau of Standards (NBS) has released several Federal Information Processing Standards (FIPS) publications regarding security of machine-readable media. Data resource management is being specifically addressed by FIPS Task Group 17.

Computers, however, also provide new capacities for protecting individual data. Once the data are in a machine-readable medium, individual and geographic identifiers can be removed from the records. The data can then be made available to interested parties with the confidentiality of the data safeguarded. Data can be encrypted much more easily in a computer system than by hand. This provides more protection to personal privacy and data confidentiality than the manual systems of yesteryear.

Much has been written during the past few years about security, privacy and confidentiality relating to records in machine-readable media. This paper attempts to summarize what has been said on the various disclosure-avoidance techniques as they relate to the machine-readable media.

Terminology

A computer system is defined broadly here to include the computer and all its peripheral devices, e.g., tapes, disks, terminals.

Confidentiality, privacy and security have been used in various ways. However, these terms will be used here in the same manner as they are presented by NBS (1973). Confidentiality is a concept which applies to data, while privacy applies to individuals. Security is the protection of hardware, software and data through the imposition of appropriate safeguards.

Restricted files are defined as those files with access controlled and limited to specific individuals or systems.

Confidentiality

Data confidentiality plays an important role in many Federal agencies' data collection procedures especially when they rely on voluntary responses. Respondents are informed that their replies will remain confidential. In order to preserve this confidentiality, appropriate steps must be taken during computer processing. One approach is using privacy transformations, also referred to as encryption, with very sensitive data. Several techniques are available.

Adding "noise" or random disturbances to each individual datum is discussed by Dalenius (1976). He describes refinements to the general procedure such as ordering the data, dividing them into groups, adding noise to the data in each group, and thus minimizing the error introduced to the group totals.

Dalenius also discussed reversible and non-reversible privacy transformations. Three methods of reversible privacy transformations discussed are Boolean addition of the key to the data, addition mod$_2$, and comparing the data to a key. These use the OR, exclusive OR, and coincidence (complement of the exclusive OR) operators respectively. All techniques are relatively easy to automate and have the desirable features of an encryption algorithm. That is, the algorithm can be known, but only those who know the key can access the data. Still, anyone with enough time, money and computer power can reverse (decode) the encrypted data. Therefore, the key must be large enough to create a very low likelihood of reversing the encrypted data. Most intruders will then be sufficiently discouraged, since the costs of exposure would likely be greater than the data are worth.

A simple example of a reversible privacy transformation is adding two keys independently to the datum being protected. Let Y be the data to be protected, K_1 and K_2 are two independent keys. Then

$$\begin{array}{r} Y \\ +K_1 \\ \hline Y + K_1 \end{array} \quad \text{and} \quad \begin{array}{r} Y+K_1 \\ +K_2 \\ \hline Y + K_1 + K_2 = X \end{array}$$

To obtain our original datum, Y, from X we must subtract the keys in reverse order, i.e.,

$$\begin{array}{r} Y+K_1+K_2 \\ -K_2 \\ \hline Y + K_1 \end{array} \quad \text{and} \quad \begin{array}{r} Y+K_1 \\ -K_1 \\ \hline Y \end{array}$$

Two non-reversible privacy transformations are discussed by Dalenius (1976) as they relate to statistical information systems. Statistical information systems differ from administrative information systems in that they serve as the basis for actions directed at *groups* of individuals or objects instead of each individual or object. Adding noise to the data is an example of a non-reversible privacy transformation. An analogy to the randomized-response design using two sets of original data is discussed. Coded data are generated according to the rules applied.

NBS (1977) published a data encrption standard which "is designed to encipher and decipher blocks of 64 bits under control of a 64 bit key." This algorithm would be built into the computer hardware, not into a computer program. Many feel that this encryption algorithm is, for all practical purposes, immune from being broken.

Security

Physical security, computer operating systems security, and file security all require attention to properly protect restricted files. These three topics are discussed below.

Physical Security

Restricted files require various degrees of physical security for obvious reasons. Natural disasters (i.e. earthquake, flood, hurricane, tornado, wind storm) fire, power failure, environmental dangers, and protection from theft, fraud and vandalism are major considerations of physical security. FIPS PUB 31 *Guidelines for Automatic Data Processing Physical Security and Risk Management*, released by NBS (1974), provides excellent physical security guidelines. Physical security procedures for restricted files include restricting access to the computer-room backup files, and storage of documentation under lock and key. They range from having security guards to simply keeping file cabinets locked. The concentric circle approach to physical security, i.e., locked files rooms, and building, provides several levels of physical security. That is, all personnel able to unlock the building door (or show proper identification to a security guard) must have a key to the room they wish to access, etc. The degree of physical security should be based on the relative value top management places on the data.

Computer Operating Systems Security

No computer operating system is completely secure. However, computer operating systems are more complex than they were twenty years ago, and for most computer systems, only an experienced systems programmer would be capable of accessing restricted files which have some degree of protection. A recent publication entitled *Security Analysis and Enhancements of Computer Operating Systems* (NBS, 1976) reviews three commercial operating systems and suggests security enhancements. The International Business Machines Corporation (1974) published a six-volume report entitled *Data Security and Data Processing* which reported results of its data security studies at four data-processing centers. These studies investigated the economic and procedural factors involved in using a secure system to determine whether and to what extent the degree of data security of a computing system can be measured, how a system can best authorize access to data, and the impact on existing systems of converting to a secure system. Most studies tend to show that some security benefits are derived simply from the analysis required to determine what security does exist.

Additional security in computer operating systems has definite limitations as well as advantages. A re cent article (Marsh 1976) noted that up to 20 percent can be added to overhead costs by having some security systems incorporated into the operating system. Additional security requirements increase the time needed to implement a new system. Production delays can occur due to the additional time required by the security system being used. Turn and Ware (1975) discuss the principles of least privilege and defensive design, which they describe as basic to any computer security system. The principle of least privilege involves limiting the user or system to those accesses and privileges needed to perform their functions, i.e., on a "need-to-know" basis. An example of defensive design is compartmentation of the system to limit the damage an intruder can do if he does succeed in penetrating a part of the protected system.

File Security

There are several techniques which can be used to secure restricted data files. Separating identifiers, e.g., name and address, from the rest of the file can provide a source of security.

Only a limited number of personnel with a "need-to-know" should have access to restricted data files. Data integrity plays a key role in this approach. The more personnel use a restricted file, the more danger there is to the security of that file.

Source data that are sensitive may be marked with appropriate classifications, i.e., SECRET, CONFIDENTIAL, etc. These classifications should be defined and justified in agency policy statements and fully explained to all personnel.

It should be clearly understood by the staff that appropriate disciplinary action will be taken if disclosure occurs. These policies should be routinely audited for compliance.

When magnetic tapes and disk packs containing non-critical sensitive data are no longer needed, they should be "erased." This can be accomplished with a computer program writing all zeroes, blanks or any other character over the sensitive data on the tape or disk pack. "Degaussing" equipment, which "scrambles" the magnetic bits on tape or disk, may be used if many tapes or disk packs require this on a recurring basis.

Passwords can be assigned to files and, depending on the computer operating system and programming language used, can often be assigned at the data-element level as well. Therefore, sensitive data elements in a file can be made secure without removing them from the file. Many, but not all, data-base management systems have this capability which is essential in a data-base environment with restricted (sensitive) data.

Data dictionary/directories (also referred to as data element directories, or data resource directories) are being used with large data bases. They can be used to check all users of the data base, i.e., what data elements and/or records the users may access, or how the data may be accessed (read, update, delete). Such a scheme could maintain a record of what users have accessed.

Programs processing restricted data files should blank out any work areas (including input and output buffer areas) in the program after the restricted file has been processed. Many computer operating systems do not clear the memory or "core" area after each job or before another job uses the space. For example, if the next job is abnormally terminated, and it is using the space where the restricted file was being processed, portions of the restricted file may be printed out and seen by the unauthorized user.

The misuse of name and address files can be monitored by including false names with valid addresses. Contacts to such names would reveal unauthorized access to the file. This technique creates some minor problems if it is used in a file where probability sampling is used, but these problems can be solved with little difficulty.

APPENDIX C

Selected Methodological Issues in Statistical Disclosure Avoidance

Dr. Lawrence H. Cox, U.S. Bureau of the Census

In this appendix, we briefly discuss the major methodological problems in the development of an automated system to perform *statistical disclosure avoidance*[1] in a *publication hierarchy*[2] such as a census or major survey. To a great extent, this description is a recapitulation of U.S. Bureau of the Census previous and current experience in developing a disclosure-avoidance system for the 1977 Economic Censuses of Wholesale and Retail Trade, Service Industries, Manufactures, Minerals and Construction Industries.

The objective of statistical disclosure avoidance is to afford confidentiality to each respondent by preventing the identification of an individual reporting unit through determination of its individual responses, and, *vice versa,* determining particular responses of a known reporting unit through other responses and published aggregates.

As a statistical procedure, this amounts to protecting the values of certain statistical cells, called the *sensitive* cells, from discovery or unacceptably narrow estimation, the process being referred to as *maintaining statistical confidentiality.* The fundamental object in this analysis is the sensitive cell and the initial problem is that of establishing a definition of sensitive statistical cell which is sufficiently broad to identify all cells involving a potential breach of statistical confidentiality, yet restrictive enough to exclude cells which are clearly non-confidential. In the the case of frequency count data, such as demographic cross-tabulations, the U.S. Bureau of the Census defines a sensitive cell by means of a *threshold rule:* if the cell contains fewer than a prescribed (threshold) number of respondents, then the cell is sensitive, and is nonsensitive otherwise. For aggregate data, such as total sales over all establishments responding in a cell, a *dominance rule* is employed: regardless of the number of respondents in the cell, if a small number (n or fewer) of these respondents contribute a large percentage ($k\%$ or greater) of the total cell value, then the so-called *n-respondent, k% rule* of cell dominance defines this cell as sensitive, since it is likely that the value of the response of one or more of these n dominating respondents or the remaining respondents may be discovered or closely estimated from the total cell value by another respondent or knowledgeable party. The value of a sensitive cell must be *masked* or protected. This may be accomplished through techniques of cell value equivocation such as rounding or perturbing cell values or by suppressing the sensitive cell and certain additional cells from publication. The latter method, known as *cell suppression,* is the Census Bureau's preferred technique in the case of aggregate data from business establishments.

Having defined the notion of sensitive cell, the statistics disseminator must define what in general is an acceptable level of protection of the value of a sensitive cell. This definition must reflect accepted methodology and procedure and, perhaps most important, must be consistent with the established definition of sensitive cell. To establish this definition, for each sensitive cell X and its corresponding value V(X), the statistics disseminator must compute two real numbers L(X) and U(X) in terms of which an interval estimate of V(X) which strictly contains the interval $L(X) \leq V(X) \leq U(X)$ is defined as an *acceptable interval estimate,* and an interval estimate of V(X) which overlaps this interval or is contained within this interval is by definition *unacceptable.* We refer to L(X) and U(X) as the *bounds of equivocation* of V(X). The entire collection of publication cells must

[1] The term *statistical disclosure avoidance* denotes both the identification of potential disclosures and the application of appropriate techniques to avoid disclosure. Some authors in the field employ the term *statistical disclosure analysis* in this regard.

[2] A *publication hierarchy* (or *publication network*) is defined as the collection of tabular arrays which constitute a statistical publication together with the linear relations between individual cell values in and across these tabular arrays.

now be analyzed as a hierarchy or network, since the linear relationships between the cell values imposed by publishing cells at different levels of aggregation are the principal means by which users of the statistics may obtain estimates of the values of sensitive cells. For example, if the value of a statistic at the State level is published and the values of the statistic at the county level are published for some but not all of the counties within the State, then the value of the statistic V(C) for any unpublished county *C* may be estimated O $\leq$V(C) $\leq$D, where *D* is equal to the difference between the value of the statistic at the State level and the sum of the *published* county values. It is in terms of these linear relationships that the statistics disseminator must develop and evaluate appropriate cell protection mechanisms. Moreover, as sensitive cells proliferate downward from one level in the hierarchy to the next lower levels (e.g., cell dominance in a cell at the State level implies that cell dominance exists in at least one of the constituent cells at the county level), then techniques of statistical disclosure avoidance in a publication hierarchy must also proceed "top-down" through the hierarchy to insure consistency of estimates from level to level and a relative minimum of disclosure processing.

The display of the cell data in published tabular form generally reflects some but not all of the linear relationships between the cells, so that these tabular displays are frequently not the actual *logical tables* upon which disclosure-avoidance techniques must be performed. For example, the Census of Retail Trade contains, for each State, a table consisting of the total sales for all establishments and the total sales for all establishments with payroll for certain retail industry classifications and their subclassifications, together with other aggregate statistical data for these industries. This set of tables represents a multiplicity of logical tables, each of which is either two or three dimensional. In particular, to each industry classification and its immediate subclassifications (immediate disaggregates), there corresponds a three-dimensional table of sales in these industries by State and by establishments with and without payroll (the latter determined by subtracting sales for establishments with payroll from total sales). This table represents a three-way disaggregation of the U.S. total of sales by industry, by State, and by payroll classification.

A disclosure avoidance system therefore cannot, in general, operate simply on the publication tables as they are displayed, but must construct all logical tables in the hierarchy and analyze these in a proper "top-down" sequence. This is a matter of identifying every level of aggregation in the publication hierarchy, appropriately sequencing these, and applying effective intra-table disclosure-analysis techniques to each logical table in turn. The suppression information is carried forward to tables lower in the aggregation hierarchy, where the internal cells of the original logical table appear as marginal totals (such as a State total being carried forward to a table of constituent counties).

A methodologically sound technique for intra-table disclosure avoidance must be applied to each logical table in turn. For the aggregate economic data to be published for the 1977 Economic Censuses cell suppression techniques will be employed. Each sensitive cell is suppressed from publication, together with as few additional (*complementary*) cells as possible to guarantee that linear estimates of the values of suppressed sensitive cells derived from the publication (such as the difference between a row of column marginal total and the sum of all published cells on the line) are acceptable estimates. Optimal suppression algorithms for two-dimensional logical tables developed at the Census Bureau will be employed in the analysis of the 1977 Economic Censuses, as well as three-dimensional suppression and analytical routines, both in tandem with linear estimation techniques designed to produce the best possible linear estimates of suppressed sensitive cells.[3] The goal is for the Census Bureau to develop and employ complete information about the disclosure potentialities contained within its publications, as can be deduced from these publications, and to be confident that only acceptable estimates of its sensitive cells can be made on the basis of the published data.

[3] For details on the optimal two-dimensional suppression strategy, see Cox (1975:380–382) and for a discussion of the improved technique of linear estimation, see Cox (1976).

APPENDIX D

Bibliography

American Statistical Association

1977—"Report of Ad Hoc Committee on Privacy and Confidentiality." *The American Statistician* (May):59–78.

Barabba, V. P. and Kaplan, D. L.

1975—"U.S. Census Bureau Statistical Techniques to Prevent Disclosure—The Right to Privacy vs. the Need to Know." Paper read at the 40th session of the International Statistical Institute, Warsaw.

Clayton, C. A. and Poole, W. K

1976—"Use of Randomized Response Techniques in Maintaining Confidentiality of Data." Draft Report. Research Triangle Institute.

Cox, L. H.

1975—"Disclosure analysis and cell suppression," American Statistical Association. Proceedings of the Social Statistics Section:380–382.

1976—*Statistical Disclosure in Publication Hierarchies*. Report No. 14 of the research project Confidentiality in Surveys. Department of Statistics, University of Stockholm, Stockholm.

Dalenius, T.

1974—"The invasion of privacy problem and statistics production—an overview." Statistisk Tidskrift, National Central Bureau of Statistics:213–225, Stockholm.

1976—*Privacy Transformations for Statistical Information Systems*. Report No. 9 of the research project Confidentiality in Surveys. Department of Statistics, University of Stockholm.

1977—*Towards a Methodology for Statistical Disclosure Control*. Report No. 19 of the research project Confidentiality in Surveys. Department of Statistics, University of Stockholm.

Fellegi, I. P.

1972—"On the question of statistical confidentiality." Journal of the American Statistical Association, 67:7–18.

1975—"Controlled random rounding." *Survey Methodology*, 1:123–133. Statistics Canada, Ottawa.

Goldfield, E. D., Turner, A. G., Cowan, C. D., and Scott, J. E.

1977—"Privacy and confidentiality as factors in survey response." Paper presented at the annual meeting of the American Statistical Association, Chicago, 1977.

Hansen, M. H.

1971—"Insuring confidentiality of individual records in data storage and retrieval for statistical purposes." AFIPS Conference Proceedings, 39:579–585. Fall Joint Computer Conference. AFIPS Press, Montvale, N.J.

Hirsch, P.

1967—"The punch card snoopers." *The Nation* (October):369–372.

IBM Corporation

1974—*Data Security and Data Processing: Study of Specific Aspects of Data Security*. IBM Corporation, White Plains, New York.

Law Enforcement Assistance Administration

1976—Regulations. Title 28. Judicial Administration. Chapter 1—Department of Justice, Part 22—"Confidentiality of identifiable research and statistical information." *Federal Register*, (December 15):54846–8.

1977—"LEAA statement on the impact of disclosure avoidance techniques". Statement prepared by Lehnen, R. G., and Eldreth, J. and submitted to the Subcommittee on Disclosure-Avoidance Techniques, Federal Committee on Statistical Methodology (November), Washington.

Marsh, R.
1976—"Making data more secure." Datamation. (October) 67–69.

Miller, A. R.
1971—*The Assault on Privacy—Computers, Data Banks, and Dossiers.* Ann Arbor: University of Michigan Press.

Murphy, M.
Unpublished—Confidentiality and its Effects on Census Data. Census Division, Statistics Canada.

Nargundkar, M. S., and Saveland, W.
1972—"Random rounding to prevent statistical disclosure." American Statistical Association. Proceedings of the Social Statistics Section, 382–385. Washington, D.C.

National Center for Health Statistics
1976—*Standardized Micro-Data Tape Transcripts.* Washington: U.S. Government Printing Office.

Newman, D.
1975—Techniques for ensuring the confidentiality of census information in Great Britain. Paper read at the 2nd session of the International Association of Survey Statisticians, Warsaw, 1975.

President's Commission on Federal Statistics
1971—*Federal Statistics.* 2 Vols. Washington: U.S. Government Printing Office.

Privacy Protection Study Commission
1976—"Notice of hearings and draft recommendations, research and statistics." *Federal Register,* Vol. 41, No. 243 (December).

1977—*Personal Privacy in an Information Society: The Report of the Privacy Protection Study Commission.* Government Printing Office, Washington, D.C.

Public Law 90–620
1968—The Federal Reports Act. 44 U.S.C. 421 (October)

Public Law 93–353
1974—The Health Services Research, Health Statistics, and Medical Libraries Act of 1974.

Public Law 93–502
1974—The Freedom of Information Act. 5 U.S.C. 552 (November).

Public Law 93–579
1974—Privacy Act of 1974, 5 U.S.C. 552a (December).

Public Law 94–409
1976—Government in the Sunshine Act, 5 U.S.C. 552b (September).

Public Law 94–455
1976—Tax Reform Act of 1976 (October).

Singer, E.
1977—"Informed consent: consequences for response rate and response quality in social surveys." Paper presented at the annual meeting of the American Sociological Association, Chicago, 1977.

Social Security Administration
1977—"Guidelines for preventing disclosure in tabulations of program data." Memorandum (February) to Office of Research and Statistics Executive Staff. Social Security Administration, Office of Research and Statistics, Washington, D.C.

Sweden, National Central Bureau of Statistics
1974—Confidentiality in Statistical Tables. National Central Bureau of Statistics, Stockholm.

Turn, R. and Ware, W. H.
1975—"Privacy and security in computer systems." *American Scientist,* 63:196–203.

U.S. Bureau of the Census
1975—Minutes of the Census Advisory Committee of the American Statistical Association, September 18–19, 1975. Washington, D.C.

1977—Reference Manual on Population and Housing Statistics from the Census Bureau, (March). Census Bureau, Washington, D.C.

U.S. Department of Health, Education, and Welfare
1973—*Records, Computers, and the Rights of Citizens.* Report of the Secretary's Advisory Committee on Automated Personal Data Systems. Washington: U.S. Government Printing Office.

U.S. National Bureau of Standards
1973—Controlled Accessibility Bibliography. NBS Technical Note 780 (June).

1974—"Guidelines for automatic data processing, physical security and risks management." Federal Information Processi .g

Standards (FIPS PUB.) 31 (June). Washington, U.S. Government Printing Office.

1976—"Security analysis and enhancements of computer operating systems," National Bureau of Standards Interagency Report 76–1041 (April).

1977—"Data encryption standard." Federal Information Processing Standards Publication (FIPS PUB.) 46 (January).

U.S. Office of Management and Budget, Statistical Policy Division

1975—"Privacy Act implementation. Guidelines and responsibilities." *Federal Register,* Part III, Vol. 40, No. 132: (July) 28948–28978. Government Printing Office, Washington, D.C.

1977—"Confidentiality issues: disclosure avoidance," *Statistical Reporter* (January): 137–138.

For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington, D.C. 20402

Stock No. 003-005-00177-8

☆ U.S. GOVERNMENT PRINTING OFFICE : 1978 O—261-071