Research Agenda

Although much research and development on disclosure limitation have been done and should be disseminated, many topics worthy of research remain. The Subcommittee has focused on fourteen topics broadly useful to Federal agencies. The Subcommittee organized these topics into three **research areas** (microdata, tabular data, and other data products) and associated with each topic a development activity designed to facilitate implementation and use.

Each Subcommittee member prioritized the fourteen research topics, and we combined these rankings to achieve the prioritization seen in Table 1 at the end of this chapter. The members' rankings varied considerably. This is not surprising. Various agencies will have different priorities because of the different types of data products they release and because of differences in the methodologies and technologies they currently use. These differences even occur within agencies causing people from different areas within an agency to have different priorities. Also, users of the data may rank these research topics differently than would suppliers of the data.

Table 1 lists only the three topics with the highest priority. In combining our rankings, we found that all Subcommittee members feel that these three topics are of great importance. The rankings of the other eleven topics listed in Table 2 were consistently lower. In general, most Subcommittee members feel that good methodology already exists for tabular data while many decisions concerning the disclosure limitation of microdata are based solely on precedents and judgement calls. Thus the Subcommittee feels that research concerning the disclosure limitation of microdata takes priority over research concerning tabular data. Also, Subcommittee members feel that research focusing on the effects of disclosure limitation on data quality and usefulness for both microdata and tabular data is of great importance.

A. Microdata

While the disclosure risk of microdata is higher than that of tabular data, the usefulness of microdata is also correspondingly higher (Cox and Zayatz, 1993). The following research topics are aimed at increasing the ability of statistical agencies to release microdata subject to confidentiality constraints.

A.1. Defining Disclosure

Primary disclosure rules for statistical tabulations are fairly well established. New research is not a high priority, particularly if agencies follow the Recommended Practices in Chapter VI. However, the problem of defining disclosure in microdata is far from solved.

The definition and the assessment of disclosure risk in microdata need to be put on a sound statistical footing. Probability theory provides an intuitively appealing framework for defining

disclosure in microdata in which we relate disclosure to the probability of reidentification. On this basis, we could measure disclosure limitation quantitatively and we could easily incorporate extensions, such as analysis based on prior knowledge. Without a measure of disclosure risk, decisions concerning the disclosure limitation of microdata files must be based on precedents and judgement calls. Research into probability-based definitions of disclosure in microdata should have high priority.

One part of this research involves developing a method of estimating the percent of records on a sample microdata file that represent unique persons or establishments in the population (Zayatz, 1991b). Another part involves developing a measure of marginal disclosure risk for each variable on a microdata file and analyzing changes in overall risk that result from changes in detail of each variable. After a measure of risk is developed, agencies may choose different maximum allowable levels of disclosure risk for different public-use microdata files.

A.2. Effects of Disclosure Limitation on Data Quality and Usefulness

A.2.a. Disturbing Data

Due to advances in computer technology and an increase in the number of available, potentially linkable data files, it may become necessary to disturb microdata prior to release in order to make matching more difficult. Some agencies have used disturbance techniques (see Section V.B.3) such as addition of random noise or data swapping, and more research is needed to investigate the protection provided by the various disturbance techniques and the usefulness of the resulting microdata (Spruill, 1983).

A.2.b. More Information about Recoded Values

Users should be consulted as to the benefit of releasing means, medians, and/or variances of all values that have been top-coded or bottom-coded and of all values in each category of a recoded variable. A minimum size requirement for each category would be necessary.

A.3. Reidentification Issues

The principal risk in releasing microdata is that a third party could match microrecords to another file containing identifiable information with reasonable accuracy. However, due to differences in sample, in responses (reporting variability), in age of data, in edit and imputation techniques, in nonresponse, and in definition and recoding of variables, linking records from two different files may be difficult. Agencies should conduct realistic attempts to match files with overlapping information, **reidentification experiments**, in order to better understand and reduce disclosure risk.

A controversial research proposal involves hiring an "intruder" to attempt to link respondents to their corresponding records on a microdata file. It would be useful to see how an intruder might approach the problem, whether or not any correct matches were made, and if correct matches were made, the amount of time and work that were required. A potential problem with

this research activity could arise if the intruder was successful in making one or more correct matches. Even if the hired intruder was an agency employee, people outside the agency could find out the results of this research under the Freedom of Information Act. This could harm the agency's reputation for maintaining confidentiality.

A.4. Economic Microdata

The feasibility of releasing microdata from economic censuses and surveys should be further investigated. Models for release or administrative alternatives need to be proposed (McGuckin and Nguyen, 1990; McGuckin, 1992). Some agencies have released very limited establishment-based microdata files, for example the 1987 Census of Agriculture files released by the Bureau of the Census. However, as stated in Chapter V, the disclosure risk for establishment microdata files is much greater than for demographic files, and data content of currently released establishment-based files is so limited that many users consider them useless.

A.5. Longitudinal Microdata

Longitudinal information increases both the identifiability of individual respondents and the amount of confidential information at risk of disclosure (see Section V.C.4.b). When linkable files are released on a flow basis, the disclosure risk of a given file depends on what information was released in previous files. The advice given in Chapter V was to plan ahead, considering disclosure limitation techniques that will be used for future files, but this is difficult because the choice of variables and their detail and variable sensitivity may change over time. Research is needed to assess the level of disclosure risk in longitudinal microdata files and to develop appropriate disclosure limitation policies and procedures.

A.6. Contextual Variable Data

Social scientists are interested in obtaining microdata files with contextual variables. **Contextual variables** are variables that describe the area in which a respondent resides (such as average income of all residents in the county) but do not identify that area. In general, the areas described are much smaller than areas explicitly identified on microdata files. Contextual variables are often costly to compute, and they can increase the disclosure risk of a microdata file because a detailed description of an area can lead to the identification of that area. Identification of such small areas is undesirable in terms of disclosure risk. Further study is needed which will identify an affordable method of generating contextual variables that will not lead to the identification of small areas (Saalfeld, Zayatz, and Hoel, 1992).

A.7. Implementation Issues for Microdata

Standard software for disclosure limitation in microdata would be of considerable benefit within and across agencies, and would serve as a quality assurance tool for Review Panels (Cox and Zayatz, 1993). This software could perform disclosure limitation techniques such as top-coding, recoding, and adding noise and could provide Panels with distributions and cross tabulations for review. Systematic development of this software should facilitate further research on improved

disclosure limitation methods for microdata and analysis of their effects on data quality and utility.

A potentially useful tool for this research and development is **matrix masking**--the representation of disclosure limitation methods in terms of matrix algebra to facilitate their implementation and analysis (Cox, 1991; Cox, 1993b). Matrix masking should be explored as a format for representing, implementing and comparing microdata disclosure limitation methods.

B. Tabular Data

The research topics below are designed to improve the efficiency of disclosure limitation for tabular data in terms of the amount of work required and the amount of information sacrificed to achieve protection.

B.1. Effects of Disclosure Limitation on Data Quality and Usefulness

B.1.a. Frequency Count Data

Research should be conducted to find out which protection method data users prefer for frequency count data (see Section IV.A). The options include controlled rounding, controlled perturbation, cell suppression, and perhaps the confidentiality edit used by the Census Bureau for the 1990 Census of Population and Housing publications.

B.1.b. Magnitude Data

Agencies normally use cell suppression for protection of confidential tabular data. Agencies should find out if data users prefer the collapsing (or rolling up) of categories to cell suppression.

Users should be consulted as to the benefit of publishing ranges for suppressed cells. Some publishing of ranges is already being done. For example, the Census Bureau publishes feasibility ranges for suppressed cells containing the number of employees in its County Business Patterns publications. These ranges, however, are larger than the feasibility ranges of those same cells that a data user could determine with a linear programming package. If users would like agencies to publish the smaller ranges, a feasibility study should be done to see if users could assimilate and manipulate information provided in this form and if a large volume of data would preclude this action due to the time needed for determining the ranges.

Users should also be consulted as to the benefit of releasing means, medians, and/or variances of all values in each suppressed cell in a table. A minimum number of respondents in each of these cells would be necessary.

B.2. Near-Optimal Cell Suppression in Two-Dimensional Tables

Network flow methods work well in two-dimensional tables. However, other methods also offer desirable features. It would be useful to specify the characteristics of and develop an optimal method for disclosure limitation in two-dimensional tables and, by combining desirable features of existing methods, to create a new method that improves upon each of the original methods. Such a method would be advantageous, as Federal statistical agencies analyze many thousands of tables each year. It would also improve the handling of three- and higher dimensional tables and interrelated sets of tables using methods based on combining two-dimensional procedures. The list of original methods includes network flow, general linear programming, integer programming, and neural networks. These methods are discussed in Chapter IV. The question is which method (or combination) is best with respect to well-defined criteria such as cost, computational efficiency, transportability, extension to higher dimensions, ease of implementation and maintenance, and data use.

B.3. Evaluating CONFID

Statistics Canada has a set of programs called CONFID which performs cell suppression on tabular data. CONFID has been made available to U.S. Federal statistical agencies. It would be worthwhile to extensively test and evaluate this system of programs and to compare it to the current cell suppression systems used at the Census Bureau and elsewhere.

B.4. Faster Software

Federal agencies would benefit from locating and purchasing the fastest network flow package available. Current cell suppression methodology uses network flow methodology when applying complementary suppressions (see Section IV.2.b). In addition, one proposed **filter technique** also uses network flow methodology to locate (and remove) superfluous complementary suppressions.

Agencies would also benefit from locating and purchasing the fastest linear programming package available. The network flow-plus-heuristic technique currently used at the Census Bureau to find complementary suppression patterns in three-dimensional tables yields non-optimal solutions. Any technique for finding optimal solutions to the cell suppression problem is currently impractical due to computer time constraints. However, there exists a linear programming technique which yields better solutions for three-dimensional tables than the currently used network-based procedure and which agencies could use if a faster linear programming package was available.

Auditing programs (see Section IV.B.2.a) also use linear programming packages. Auditing programs are programs that check to see if all primary suppressions in a table are indeed sufficiently protected after complementary suppressions have been applied. So, a faster linear programming package would lead to faster auditing programs.

B.5. Reducing Over-suppression

One problem with the currently used cell suppression methodology is that it applies complementary suppressions to only one primary suppression at a time. The system has no way of considering all of the primary suppressions at once, and this leads to the application of too many suppressions (**over-suppression**). There may be one or more procedures which could be applied prior to using current techniques and which would reduce the amount of over-suppression caused by this one-primary-at-a-time approach. These procedures should be developed and tested.

To reduce over-suppression, the Census Bureau is currently investigating the possibility of incorporating one small integer programming tool into the current cell suppression procedures to obtain better results. The purpose of this particular integer program is to eliminate superfluous complementary suppressions (Sullivan and Rowe, 1992). Other ways of using integer programming procedures to reduce over-suppression should be examined.

There may be one or more procedures that could be applied after the current cell suppression techniques to reduce the amount of over-suppression caused by the one-primary-at-a-time approach. These techniques, often called **filter** or **clean-up techniques**, examine the table including the complementary suppressions and attempt to locate cells which were chosen as complementary suppressions but which could be published without a loss of protection of the primary suppressions. One procedure involves using network flow methodology repetitively to locate the superfluous complementary suppressions. The practicality of using this filter technique, which increases the computer time needed to perform disclosure analysis considerably, should be investigated. Other filter techniques should be developed and tested.

Another major drawback of the currently used cell suppression methodology is that, often, due to computer storage and methodology constraints, not all data values in all additive relationships can be considered simultaneously. This is particularly true for the large amount of tabular data published by the Census Bureau. Thus the problem must be broken down into pieces (sets of data) that are processed separately. Unfortunately, some data values are necessarily in more than one piece. While trying to ensure that the various pieces, when considered as a whole, do not result in a disclosure of confidential information, it is necessary to reprocess many of the pieces several times. This reprocessing of sets of data due to the inability to process all data simultaneously is called **backtracking**. Backtracking is time consuming and results in over-suppression. Research that could reduce the amount of backtracking needed and the over-suppression caused by backtracking would be beneficial.

C. Data Products Other Than Microdata and Tabular Data

Previously, disclosure limitation research focused on either microdata or tabular data. However, agencies are also releasing information in the form of database systems and analytical reports. Research is needed to analyze disclosure risk and develop and evaluate disclosure limitation techniques for these types of data products. The next section describing database systems is

lengthy because we provided no background information on this subject in previous chapters. The length of the section does not reflect the importance of the research topic.

C.1. Database Systems

More research is needed to analyze the feasibility of storing and allowing access to microdata in a database system with security controls and inferential disclosure limiting techniques (Keller-McNulty and Unger, 1993). Research in database systems assumes quite a different approach than we have discussed so far. Rather than releasing known subsets of data, it is possible to keep all data on-line in a **database management system** (DBMS) that dynamically enforces the controls. Normally this is a relational database that organizes the data in a series of tables. These data tables are similar to microdata files, although they might also contain sensitive data such as salaries, which are releasable only as an aggregated value such as average salary. Users may request whatever subset of information they need, and the database may return either an exact or approximate answer, or even refuse to answer if the data would disclose individual identities. The relational database management system, therefore, simplifies retrieving data, but does not actually improve availability, i.e., exactly the same data are still available. References considering such an approach to statistical databases are (Adam and Wortman, 1989) and (Michalewicz, 1991).

If such a system were successful, it would allow the greatest possible use of the data, while still maintaining the confidentiality of the individual. Such a system could allow for use of microdata containing administrative data. The system would handle special requests without any manual intervention. Reports, special files, tapes, etc. would not be generated unless actually requested. Present research in this type of system has discovered many problems, and proposed a few solutions, but no one technique is presently accepted as a standard. Implicit in such proposals is the requirement that the data can only be accessed via the database, not directly through the file system or by transmission protocols. Encryption of the data by the database is one method of protecting the data during either storage or transmission.

Implementation of dynamic controls on data access could possibly be accomplished using some developing concepts from the area of database security. For this purpose, we can assume that some information is more sensitive than other information. The sensitive information is classified "High", while the less sensitive information is classified "Low". For example, the entire survey could be available on-line, but the individual's name (and other identifying information) could be classified as "High". Names would be available only to survey personnel. Such a database would have to incorporate various protection mechanisms common to secure database systems. Of most significance to statistical database management systems would be the assurance that no process or subroutine has been surreptitiously inserted into the system to simply write protected information into a "Low" area (such a process is called a "Trojan Horse"). If users are limited to reading data then the disclosure problem becomes identical to the **inference problem** of interest to database security researchers. Can it be guaranteed that "Low" users cannot infer the information from the other, authorized, Low data that they retrieve? This is an open problem that has been solved only for a few specialized types of inferences. Perhaps the most promising approach is to define a set of constraints, or rules, to be checked for each query.

Such constraints could enforce query set size, overlap, or complexity restrictions. Constraints could dynamically simulate cell suppression techniques. They could also correlate recent queries to insure that the particular user cannot combine these answers to infer unauthorized information. Research into such inference issues would be especially valuable since it addresses the inference links that exist between separate tables (or relations), or even users. It, therefore, addresses the question of whether the data in one table can be used to compromise the data in another table. This is an increasingly common issue due to the growing popularity of relational databases. A good reference for inference control is (Qian, Stickel, Karp, Lunt, and Garvey, 1993).

The ability to calculate aggregates (such as averages, sums, counts, etc.) over sensitive data leads to many problems in maintaining secrecy of the data. Successive queries over varying subsets of data may be sufficient to determine specific values of the sensitive data. This is known as a **tracker attack** (Denning, 1982). Prevention of tracker attacks using query histories requires additional computation steps (Michalewicz, 1991), but using an algorithm that does not keep a history seriously restricts the amount of data available to the user.

These problems have taken on increased importance in recent years because of research in the field of **knowledge discovery** which concerns artificial intelligence based programs whose purpose is to discover relations between data items. Unauthorized disclosures may therefore be automated, and the probability of such disclosure is dramatically increased. These methods will change the way that disclosure risk must be calculated. Rather than selecting a target individual, the program will search the statistical database for *any* specific individual, and then match the one found to the outside database. For example, if average income is released by county, and some county has only one doctor, then that doctor's income may be determined. Although the probability of being able to determine the income for any arbitrary doctor is quite small, a knowledge discovery program will check the entire database to find the one doctor whose income can be determined. Such techniques have proven quite useful to direct marketing companies.

The database techniques currently used tend to emulate the paper world, and are adequate for a certain balance of protection and accessibility. If we wish to provide more accessibility, while maintaining the same level of protection, we will require techniques that are only available through automated computer programs. These will tend to require extensive computations. In fact, some protection proposals suggest that all possible combinations of characteristics be checked to insure that individual names cannot be determined from the data. Such computations may currently be suitable for small or simple datasets that can be verified once and then widely distributed. The cost of computing power must decline further before such methods are feasible for complicated databases, however.

The National Center for Education Statistics has released compressed data with access controls along with software that allows users to obtain weighted, minimum cell count cross tabulations of the data without being able to examine individual data records. Research should be done which investigates the disclosure risk in releasing results of other types of statistical analysis of the data.

C.2. Disclosure Risk in Analytic Reports

The Center for Economic Studies (CES) at the Census Bureau releases output from statistical models, such as econometric equations, estimated using confidential data. Some agencies, such as the Bureau of Labor Statistics, have fellowship programs and some, such as the Census Bureau, can "swear in" researchers as special employees to allow use of confidential data for analytical purposes. Fellows and other researchers would like to publicly release the results of their statistical analyses.

Often the resulting output from the statistical analyses takes the form of parameter coefficients in various types of regression equations or systems of equations. Since it is only possible to recover exact input data from a regression equation if the number of coefficients is greater than or equal to the number of observations, regression output generally poses little disclosure risk because normally the number of observations is much larger than the number of coefficients. One question to be addressed, however, is if the number of observations is only slightly larger than the number of coefficients, how closely can a user estimate the input data?

Also, sometimes researchers use dummy (0,1) variables in statistical models to capture certain effects, and these dummy variables may take on values for only a small number of observations. Currently, CES treats these dummy variables as though they were cells in a table and performs disclosure analysis on the observations for which the dummy variables take on the value of 1. CES applies the n,k rule to these "cells" based on total value of shipments in deciding whether or not to release their corresponding regression coefficients. Research is needed to determine if this technique leads to withholding too much information.

Table 1
Prioritization of the Three
Most Important Research Topics

Priority	Research Topic	Data Type
1	Defining Disclosure	Microdata
2	Data Quality and Usefulness	Microdata
3	Data Quality and Usefulness	Tabular Data

Table 2
Other Research Topics

Research Topic	Data Type
Reidentification Issues	Microdata
Economic Microdata	Microdata
Longitudinal Microdata	Microdata
Contextual Variable Microdata	Microdata
Implementation Issues	Microdata
Near-Optimal Cell Suppression in Two-Dimensional Tables	Tabular Data
Evaluating CONFID	Tabular Data
Faster Software	Tabular Data
Reducing Over-Suppression	Tabular Data
Database Systems	Other
Analytic Reports	Other