**Microdata Disclosure Protection**
**Research and Experiences at the US Census Bureau**

Sam Hawala
4700 Silver Hill Rd
Washington, DC 20233
Paper presented at the Workshop on Microdata, Stockholm Sweden
August 21, 22, 2003

Key words: Confidentiality Protection, Statistical Disclosure Control, Microdata, Uniques.

## Abstract

In this paper we share, with workshop participants, our experiences and research areas in meeting our legal and ethical requirements to avoid disclosing information provided by respondents, with a particular focus on public-use microdata. The U.S. Census Bureau methods of Statistical Disclosure Limitation reflect current internationally established practices. Research at the Census Bureau consists of refining and improving existing methods, as well as creating new methods to respond to changing data environments.

## 1. Introduction

The Census Bureau is one of more than 70 U.S. federal agencies that have a role in collecting data from individuals, households, farms, businesses, and governmental bodies and disseminating those data for a variety of statistical purposes. All of these agencies are caught in the inherent tension between data protection and data access.

The Census Bureau faces a growing challenge to meet the American public's data needs, while maintaining a commitment to protect respondent confidentiality. Advances in technology and proliferation of public and semi-public data repositories are making it increasingly more difficult to prevent re-identification of survey respondents as a possible result of data linking attempts.

The Census Bureau conducts its multipurpose statistical programs under government-wide legislation such as the Privacy Act, the Freedom of Information Act (FOIA), and the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) of 2002; and agency-specific legislation such as Title 13, United States Code, of 1954. Title 13 requires the U.S. Census Bureau to protect the confidentiality of respondent information. Those who may access confidential data essentially fall into two categories: (1) employees and (2) non-employees who receive Special Sworn Status (SSS). Title 13 permits the Census Bureau to provide SSS to non-employees who help the Census Bureau carry out its work, by making them liable for penalties for unauthorized

disclosure or use of protected information, just as employees are. Unauthorized disclosure is punishable by a fine of up to $250,000 or a jail term of up to 5 years or both.

The Census Bureau recently formed a Data Stewardship Executive Policy committee (DSEP) to assure that the Census Bureau can effectively collect and use data while fully meeting the legal and ethical obligations to respondents to respect privacy and protect confidentiality. The committee oversees the work of a Disclosure Review Board (DRB), which supports the committee by proposing policies and setting methodologies underlying confidentiality protection; reviewing external products for potential disclosure; and coordinating the confidentiality-related activities needed to inform decision-making on data collection, data linking, and data dissemination. In reporting to the DSEP, the DRB coordinates its efforts with three other committees: the Privacy Policy and Research Committee, the Committee for Administrative Records Policies and Procedures, and the Enterprise Security Issues and Policy group.

The ability of the Census Bureau to continue the release of microdata – records of information on individual respondents - that are adequately protected against disclosure is of utmost importance to its mission. Microdata is the raw material for statistical analyses of population characteristics and public policy alternatives. It is only by building and maintaining public trust and satisfying the public's needs that the Census Bureau can continue to meet its program objectives.

## 2. Data Access and Data Products

Following its mission to be the "leading source of quality data about the Nation's economy and people," the Census Bureau currently makes a great deal of data public. However, because of concerns about the possibility of a disclosure of individual information, the Census Bureau limits the amount of detailed data provided to all users in tabulations and public-use microdata files. The Census Bureau interprets Title 13, United States Code (U.S.C.), Section 9 as essentially creating two, mutually exclusive categories for Census Bureau data: confidential or publicly available. Data typically pass from the first category to the second by applying a set of disclosure techniques and tests. Some mechanisms exist to provide access to more detailed information on a restricted basis. These include on-site access at the Census Bureau, or in an approved Census Bureau site at another location for people with Special Sworn Status (SSS); and remote on-line access in state data centers and census information centers via the Advanced Query System for user-defined tables from Census 2000. The latter system allows users to request certain types of tables and then automatically reviews them to avoid disclosing confidential information. Users receive only the tables that have passed disclosure review and never access the microdata.

Every state, D.C. and Puerto Rico has a data center. These and some other 1800 agencies such as chambers of commerce, state agencies, libraries, local planning agencies, and universities provide web-oriented, dynamically generated data products with hyperlinks to data, metadata, and applications. For example, users can access profiles constructed

from integrated data sets at http://www.stats.indiana.edu/uspr/a/us_profile_frame.html. State data centers get more than a million data requests each year.

The U.S. Census Bureau conducts surveys on a reimbursable basis for other U.S. federal agencies and other organizations. When those surveys are based on a frame protected from disclosure by Title 13, the data may not be released to the sponsor until they have cleared disclosure review. Some agencies feel the need to see the underlying data, in order to better understand the results of their survey. In some cases, those agencies work directly with the program areas; in others, the sponsors may make use of the Census Bureau secure sites at headquarters or remote Research Data Centers (RDC) to access those data directly, as SSS personnel.

## 2.1 Remote Secure Sites – Research Data Centers

In 1983 the Census Bureau established the first research data center. It was set up at heardquarters and is called the Center for Economic Statistics. The expansion of the RDCs program began in 1998. Currently there are eight research data centers. They are located in Boston, Carnegie Mellon, University of California Berkeley, University of California Los Angeles, Duke University, University of Michigan, and Chicago.

One of the chief goals was to improve the programs by permitting access to nonpublic economic manufacturing census and survey microdata because of the difficulty in developing useful public-use files that could mask the identity of very large firms. Researchers submit proposals that undergo a project review process, which assesses the feasibility of the work, scientific merit, disclosure risk, and usefulness to the Census Bureau – benefit to the Census Bureau is a key criterion. Once the proposal has been accepted, the researchers undergo a security clearance and are given SSS. They are then permitted to use CES to access nonpublic Census Bureau data either at headquarters or at one of the Census Bureau's secure remote sites.

Survey sponsors can obtain access to the microdata for approved projects. Among the agencies that have such approved projects are: The Federal Research Board for the Census of Plant Capacity; The Agency for Healthcare Research and Quality for the Medical Expenditure Survey – Insurance Component; and the National Center on the Educational Quality of the Workforce for the National Employer Survey.

In rare cases, the researchers access extracts of the Business Register in microdata form. They can see such items as company name and Employer Identification Number within the confines of the secure remote site for linking purposes, either over time or across economic surveys. The linkage using these variables can be carried out only under the direct supervision of an on-site Census Bureau employee.

## 2.2 Online Access

The Census Bureau provides through the internet (FTP) and CD-ROM, public-use microdata products produced from the following surveys:

- Current Population Survey (CPS);
- Survey of Income and Program Participation (SIPP);
- National Health Interview Survey (NHIS);
- American Housing Survey (AHS);
- National Health and Nutrition Examination Survey (NHANES);
- Survey of Program Dynamics (SPD);
- National Ambulatory Medical Care Survey;
- National Hospital Ambulatory Medical Care Survey;
- American Community Survey (ACS); and
- Consumer Expenditure Survey.

Additionally two sets of Public Use Microdata Sample (PUMS) files, extracted from Census 2000, and are provided by state, on the main census Internet site.

The PUMS files are based on a sample of the decennial census "long form" or "sample" data. The data that the Census Bureau attempts to collect from one hundred percent of the population through what we call census "short form" questionnaires, consist of characteristics such as sex, age, Hispanic/NonHispanic, race, relationship to householder, and tenure (owner or renter.) Approximately a one in six sample of the population receives the census long form, which, in addition to the short form information, collects data on characteristics such as marital status, school attendance and grade level, ancestry, language, place of birth, citizenship, military service, income, industry, and occupation.

Two PUMS files are produced for each state: a 1-percent PUMS and a 5-percent PUMS. The 1-percent files provide the maximum amount of social, economic, and housing information available to the public. The 5-percent files offer more geographic detail but less characteristic information than the 1-percent files.

Finally, users at state data centers and census information centers can obtain online tabulations based on census 2000 data through the Advanced Query System. The Advanced Query System allows these users to select a population universe, geographic areas, and variables for tabulations that previously were not standard at the Census Bureau. Requested tables must pass an automated disclosure review using confidentiality filters before they may be sent to the users.

## 3. Microdata Disclosure Protection

### 3.1 Disclosure Threats

Society has developed an insatiable appetite for all kinds of detailed information for many worthy purposes, and modern systems tend to distribute information widely. The proliferation of public information made available on the Internet and recent access to inexpensive, fast computers with large storage capacities make today's society far more technically empowered than ever before. In the past, a person seeking to reconstruct private information was limited to visiting disparate file rooms and engaging in labor-intensive review of printed material in geographically distributed locations. Today's

ability to access voluminous worldwide public information using an inexpensive computer and easily available network resources has eroded some of the protections previously available to avoid disclosing confidential information.

An example of how technology continues to drive the desire for data is that, in the recent past, security concerns have held back some medical institutions from pursuing wireless data handling solutions because of security vulnerabilities. The Wi-Fi (short for "wireless fidelity") technology, may be gaining acceptance currently as an alternative to a wired local area network. Notwithstanding, the Health Insurance Portability and Accountability (HIPAA) Act of 1996, which requires that all data on patients be kept secure and private, some hospitals have began using Wi-Fi technology to provide bedside check-in of patients, while at the same time allowing doctors to access that same information back in their offices and exam rooms. This use of technology may not have direct bearing on how the Census Bureau disseminates data, but it supports the argument that there is, everywhere, a strong tug and pull towards efficient and easier access to data, eroding previous confidentiality safeguards.

Some federal agencies conduct record linkage exercises involving person-specific data for several reasons including their own research and for statistical information to use in building frames for future surveys. Private companies do the same thing to build more complete customer databases. There is a danger that an organization may attempt to use the linkage software and other data sets to attach identifiers to the Census Bureau public use files. The Census Bureau must ensure that this cannot be done. This challenge is particularly severe when panel data are involved, especially when they have been linked to administrative data, or when information about employees and their employers are combined.

### 3.2 Disclosure Protection Techniques

As noted earlier, those who may access Title 13 confidential data for approved work fall into two categories: (1) employees and (2) non-employees who receive Special Sworn Status (SSS). Non-employee access to Census confidential data takes place at Census Bureau facilities or occasionally at a Census Bureau-approved secure location.

In order for an individual to qualify for SSS, the project must benefit the Census Bureau’s programs that rely on data protected by Title 13. The project must be viable, with scientific merit, and without disclosure risk. In addition, individuals and their organizations must not be subject to any oath or pledge that conflicts with the Title 13 pledge of confidentiality. They must have a good track record for handling sensitive or confidential data.

When other agencies commission the Census Bureau to conduct surveys under the authority of Title 15, U.S.C., they provide the frame for the sample and the resulting data belong to the agency, subject to certain constraints. In particular, the funding agency must agree not to permit re-disclosure of the data in a manner that would permit identification

of the data of an individual respondent, and the other agency is responsible for protecting the data against disclosure.

Other surveys and censuses are conducted under Title 13, where the Census Bureau is responsible for protecting the information. The Disclosure Review Board (DRB) at the Census Bureau reviews all Title 13 disclosure-limited data products before they are released. One tool that facilitates this review is the *"Checklist on Disclosure Potential of Proposed Data Releases"* called *The Checklist.* It can be found at http://www.census.gov/srd/sdc/. It is a series of questions that are designed to assist the DRB in determining the suitability of releasing either public-use microdata files or tables. Section 3 of the Checklist pertains to microdata files. The Checklist focuses on major areas such as geographic information, variables presenting unusual risk of individual disclosure, contextual or ecological variables, possible links to external data, and possible cross-tabulations that might identify a unique combination of attributes.

Methods of protecting microdata files include combinations of the following:

- Removal of direct identifiers such as names, addresses, telephone numbers, social security numbers, or establishment identification numbers.
- Setting geographic population thresholds:
  Some surveys do not publish geographic identifiers that identify areas smaller than 100,000 people (CPS, AHS). Others use higher limits (SIPP uses 250,000); still others identify only U.S. regions (NCVS).
  The 5-percent PUMS provide information for Public Use Microdata Areas (PUMAs) with a minimum population threshold of 100,000. On the 1-percent PUMS, the areas identified are called super-PUMAs. They have a minimum population threshold of 400,000.
- Data swapping:
  In the PUMS files a percentage of households were swapped. The swapped households had to share a few characteristics while residing in different geographic locations.
- Global recoding:
  On the 5-percent PUMS files, categories of variables with less than 10,000 persons nationally are collapsed into more general categories. On the 1 percent national PUMS file this threshold is dropped down to 8,000 for the identification of categories in the race and Hispanic origin variables. Industry and occupation codes are collapsed into broader categories.
- Rounding:
  Values of income, utility costs, mortgage payments, hazard insurance premiums, rent, and other costs are rounded.

- Top-coding:
  Income variables are rounded and top-coded independently. The average values of cases at and above the top-code by state are substituted for incomes at or above the top-code values.
- Age detail
  Age is usually top-coded at 90, although each state PUMS file has the average age of individuals in the state 90 years and over.

The Census Bureau also protects confidentiality by considering other issues. For example, sampling information (identity of PSU's or control numbers) has to be scrambled; and if some data items isolate readily identifiable subpopulations, then blanking, imputation, recoding, or noise inoculation are used.

The Census Bureau also conducts research to find members of the sample who are unique, with respect to sets of key variables. If some are found, they may indicate which variables need noise inoculation. Those that remain unique when geographical detail is reduced are more likely to be unique within the entire population, with respect to the same key variables. Those cases are usually suppressed. An example is a case of a very young college graduate.

Many of our demographic public use data files are hierarchical in nature. The hierarchy is due to the fact that an id number links records of the people comprising a household. Households may be unique because of the type of association of the individuals comprising the household, through birth or marriage, or through divorce. Individuals, by themselves, may not stand out, but through association with some other individuals in the household, they may make the household unique even when the geographic detail is greatly reduced. The Census Bureau looks for variables that make a household unique and uses noise addition, local suppression, or other means to protect confidentiality.

### 3.3 Synthetic data

The idea of providing entirely simulated datasets without any disclosure risk, in place of the corresponding original datasets, has great appeal. Rubin (1993) proposed the use of multiple imputations for data simulation as a way of limiting disclosure. The simulated data must reproduce a wide variety of summaries of the original distributions if the data are to be useful to a broad range of users. The Census Bureau is in the process of developing prototypes for building synthetic data.

There are several approaches to constructing synthetic data. The Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) program is trying to build on the success of the Federal Reserve Imputation Technique Zeta (FRITZ) algorithm to mask longitudinal linked data. The FRITZ algorithm, described in Kennickell (1997), produces synthetic data for only a subset of cases and variables, leaving the remaining data unmasked. Reiter (2002, 2003) is investigating the feasibility of constructing partially synthetic public use data sets out of collected CPS data. Reiter mentions that specifying imputation models for survey data is a difficult task. Typically surveys collect data on

hundreds of variables whose distributions are not easily modeled with standard parametric tools.

3.4 Re-identification experiments

Re-identification experiments can shed additional light on the particularities of a microdata set. Hence, before the Census Bureau releases a microdata set, the Disclosure Review Board may decide to consider some additional information on the nature of the data file. The information includes:

- The number and distribution of records found to be unique;
- The amount of error in the data;
- The availability of external files with comparable data content. All forms of public or proprietary external files are considered: other microdata files, macrodata files (or tabular data), and databases allowing queries of microdata records.
- The resources that may be needed by someone attempting to identify individual units.

Experience in re-identifying respondents from de-identified microdata sets shows that the experiments should be run on a periodic basis to continually update statistical disclosure limitation strategies.

## 4. References

Abowd, M.J. and D.S. Woodcock (2001). Disclosure Limitation in Longitudinal Linked Data, in *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes (eds.), Amsterdam: Elsevier Science B. V., 215-277.

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

Cox, L. (1994). Matrix Masking Methods for Disclosure Limitation in Microdata. *Survey Methodology*, Ottawa: Statistics Canada, 20, 165-169.

Doyle, P., Lane, J.I., Theeuwes, J.J.M., and Zayatz, L.M. eds. (2001). *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, Amsterdam: Elsevier Science B. V.

Duncan, G.T., Jabine, T.B., and De Wolf, V.A. eds. (1993). Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics. Panel on Confidentiality and Data Access, National Research Council and the Social Science Research Council, National Academy Press.

Reiter, P.J. (2002). Satisfying Disclosure Restrictions with Synthetic Data Sets. *Journal of Official Statistics* 18, 531-544

Reiter, P.J. (2003). Inferences for Partially Synthetic Public Use Microdata Sets. *Survey Methodology,* forthcoming.

Rubin, D.B. (1993). Discussion of Statistical Disclosure Limitation, *Journal of Official Statistics*, vol. 9, no. 2, pp. 461-468.

Skinner, C.J., and Elliot, M.J. (2002). A Measure of Disclosure Risk for Microdata, *Journal of the Royal Statistical Society: Series B* (Statistical Methodology), 64: 855–867.

Skinner, C.J., and D.J. Holmes (1998). Estimating the Re-identification Risk Per Record in Microdata, *Journal of Official Statistics*, Stockholm: Statistics Sweden, 14: 361–372.

Willenborg, L.C.R.J. (2001). Elements of Statistical Disclosure Control, New York: Springer.

Zayatz, L. (2000). SDC in the 2000 U.S. Decennial Census. *U.S. Census Bureau Technical Report.*