Journal of Official Statistics, Vol. 20, No. 1, 2004, pp. 115–124

American FactFinder: Disclosure Limitation for the Advanced Query System

Sam Hawala[1]*, Laura Zayatz, and Sandra Rowland*

American FactFinder is the United States Census Bureau's new online data dissemination system that accesses data from the 1990 Decennial Census, the American Community Survey, the 1997 Economic Census, and Census 2000. Most of the data files accessed are summary files with matrices of aggregated data. An additional capability developed as a separate system, the Advanced Query System, is the production of tabulations from a query of the Census 2000 microdata files. The dissemination of tabulations online from a query of the microdata files requires special techniques for disclosure limitation. These techniques, described in this article, are applied to Census 2000 microdata files and the Advanced Query System.

Key words: Confidentiality; remote access; online query; microdata; U.S. Census Bureau data.

1. Introduction

Statistical agencies often collect data under a promise of confidentiality to their respondents. At the same time, they are required to publish as much high quality statistical information as possible. Therefore they must develop methods for providing access to statistical data while limiting the risk of disclosure of confidential information. Remote access is one such method (Blakemore 2001). It involves providing access to the data over secure electronic lines to dedicated computers and/or electronic access to databases previously subjected to statistical disclosure limitation techniques. The Advanced Query System is a remote access system.

The U.S. Census Bureau is the pre-eminent collector and provider of timely, relevant, and quality data about the people and economy of the United States.

In more than 100 surveys annually and three censuses a decade, evolving from the first census in 1790, the U.S. Census Bureau provides official information about America's people, businesses, industries, and institutions. The U.S. Census Bureau guarantees the confidentiality of individual responses for persons for 72 years, as required by federal law (Title 13, Section 9 of the U.S. Code). The cooperation of citizens, enterprises and other respondents in providing appropriate data needed for necessary statistical compilations largely depends on the U.S. Census Bureau achieving a balance between protecting

[1] U.S. Bureau of the Census, 4700 Silver Hill Rd., Commerce/Census/SRD/3209-4, Washington DC 20233, U.S.A. Email: Sam.Hawala@Census.Gov

Acknowledgment: This article reports the results of research and analysis undertaken by U.S. Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

individual confidentiality and allowing distribution of information in a useful and timely manner. To insure that any data accessible to external users – those outside the U.S. Census Bureau – will not disclose information on individuals or entities, the U.S. Census Bureau has developed a set of automated disclosure limitation techniques. The purpose of this article is to describe these disclosure limitation techniques for the Advanced Query System.

2. The American FactFinder System

Many of the data products released by the U.S. Census Bureau are in the form of tables that show details for Census blocks. Census blocks, averaging 34 people each, are the smallest geographic areas for which the Bureau collects and tabulates decennial census data. The data that the Census Bureau attempts to collect from one hundred percent of the population through what we call "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 long census form, which, in addition to the short form information, collects information on characteristics such as marital status, school attendance and grade level, ancestry, language, place of birth, citizenship, military service, income, industry, and occupation. The reader can find the long form questionnaire in Appendix D of the documentation at http://www.census.gov/prod/cen2000/doc/pums.pdf

The sample data are also published in the form of tables, some of which are at the block group level. More detailed tables are published at the tract level.

The average block group contains 1,348 people. Tracts are statistical subdivisions of counties delineated by local committees of census data users in accordance with U.S. Census Bureau guidelines for the purpose of collecting and presenting decennial census data. These neighborhoods contain between 1,000 and 8,000 people, or on average, typically 1,700 housing units and 4,300 people. Tracts are designed to have homogeneous population characteristics, economic status, and living conditions at the time they are established. There were more than 60,000 census tracts in year 2000.

With the proliferation of inexpensive and powerful computer systems and storage it became clear that some research goals could best be met only if the data were in microdata form; that is, disaggregated data at the respondent level. In 1963 the first public use microdata files were generated from a sample of the 1960 Decennial Census and released to the public. Public use microdata samples contain individual records of responses to questionnaires with unique identifiers (names, addresses, etc.) removed so that the confidentiality of respondents is protected.

Taking part in a government initiative aiming to make government more efficient and accessible to the public, the U.S. Census Bureau announced, in October 1998, the new Internet data-delivery system that significantly expands users' access to the agency's vast data resources. This new system complements the U.S. Census Bureau's existing Internet site by giving the public online access for the first time to the Census Bureau's largest data sets. The new system is referred to as "American FactFinder" (AFF). It was built under

contract with the U.S. Census Bureau by IBM Global Services Corp., principal contractor, responsible for systems integration and user-interface design.

The first data released via AFF were preliminary reports from the 1997 Economic Census, the 1990 Census of Population and Housing files, the American Community Survey test, and demonstration data and results of the Census 2000 Dress Rehearsal conducted in 1998. The full range of Census 2000 data products started becoming available via AFF as of January 2001, with the release of the state population totals to reapportion the U.S. House of Representatives' 435 memberships among the 50 states. The U.S. Census Bureau released the redistricting population data at the census block level on a state-by-state basis during March 2001. Redistricting is the process of revising the geographic boundaries of areas from which people elect representatives to the U.S. House of Representatives, a state legislature, a county or city council, a school board, and so forth.

The Advanced Query System allows users to select a population universe, geographic areas and variables for tabulations that previously were not standard at the U.S. Census Bureau. Some limits, described in this article, are imposed on these nonstandard tabulations. Nonstandard tabulations, in particular tabulations for small geographic areas, can introduce a risk of disclosure of individual information. Threats of disclosure of individual information may affect people's willingness to cooperate with the censuses and surveys conducted by the U.S. Census Bureau. The U.S. Census Bureau has devoted considerable staff resources to developing procedures to protect confidential data. The research and development of the disclosure limitation rules and techniques, which are presented in this article, are part of the U.S. Census Bureau's efforts in maintaining respondents' confidentiality.

3. Disclosure Limitation Rules and Techniques

Through AFF, and within the limits of the data provided, the public is able to obtain summarized data from Census 2000 over the Internet. Data are provided in table format, displayed back to the user, and in the softcopy file equivalent of the table. The U.S. Census Bureau publishes some basic tables in PDF format providing some national and state demographic profiles. These are called "Tier 1 data" products. An example of a Tier 1 table is given in Figure 1. Data products such as redistricting data used for voting purposes, as well as more than 300 other predefined table shells from the short form decennial data and over 800 tables from the long-form decennial data, from which the user can select, are referred to as "Tier 2 data." An example of a Tier 2 table is given in Figure 2. The data feeding the predefined tables are obtained from decennial summary files that are edited by the Bureau after the decennial data collection phase is complete.

When external users define their own tabulations in the Advanced Query System, the resulting tabulated output is referred to as "Tier 3 data." An example of a Tier 3 table is given in Figure 3. In Tier 3, the data feeding the tables are microdata files of individual persons and households. To insure that any data tabulation requested by external users will not disclose respondents' identities, the U.S. Census Bureau uses data recoding and data swapping (Zayatz 2003) for Tiers 1, 2 and 3 data. The U.S. Census Bureau's Disclosure Review Board has already approved all the table formats in Tier 2. Tabulated data from

Geographic area	Total households	Percent of total households							Average population per-	
		Family households				Nonfamily households				
				Type of family			Householder living alone			
		Total	With own children under 18 years	Married-couple family	Female householder, no husband present	Total	Total	65 years and over	Household	Family
United States	105,480,101	68.1	32.8	51.7	12.2	31.9	25.8	9.2	2.59	3.14
Alabama	1,737,080	70.0	32.3	52.2	14.2	30.0	26.1	9.8	2.49	3.01
Alaska	221,600	68.7	39.9	52.5	10.8	31.3	23.5	4.1	2.74	3.28
Arizona	1,901,327	67.7	32.0	51.9	11.1	32.3	24.8	8.6	2.64	3.18
Arkansas	1,042,696	70.2	32.1	54.3	12.1	29.8	25.6	10.4	2.49	2.99
California	11,502,870	68.9	35.8	51.1	12.6	31.1	23.5	7.8	2.87	3.43
Colorado	1,658,238	65.4	32.8	51.8	9.6	34.6	26.3	7.0	2.53	3.09
Connecticut	1,301,670	67.7	32.2	52.0	12.1	32.3	26.4	10.1	2.53	3.08
Delaware	298,736	68.5	31.9	51.3	13.1	31.5	25.0	9.1	2.54	3.04
District of Columbia	249,238	49.9	19.9	23.9	19.9	51.0	42.9	10.2	2.16	2.97

Fig. 1. Example of a Tier 1 data product

	Census Tract 1, District of Columbia, District of Columbia
Total:	4,674
White alone	4,192
Black or African American alone	219
American Indian and Alaska Native alone	23
Asian alone	189
Native Hawaiian and Other Pacific Islander alone	11
Some other race alone	11
Two or more races	29

U.S. Census Bureau
Census 2000

Fig. 2. Example of a Tier 2 data product

Tier 3 will be provided only if the table request passes automatic filters with the disclosure limitation rules described in this article.

3.1. Data recoding

Variables such as detailed race, age, occupation, industry, Hispanic origin, and group quarters are recoded, i.e., some sparse categories are combined to show less detail. For instance single years of age are combined into a short list: (Under 1 year, 1 year, 2 years, 3 years, 4 years, 5 years, 6 years, 7 years, 8 years, 9 years, 10 years, 11 years, 12 years, 13 years, 14 years, 15 years, 16 years, 17 years, 18 years, 19 years, 20 years, 21 years, 22 to 24 years, 25 to 29 years, 30 to 34 years, 35 to 39 years, 40 to 44 years, 45 to 49 years, 50 to

Rows: 17 Columns: 1

3-digit ZCTA (1)	5-digit ZCTA (1)	Number of Related Children Under 18 Years in Household	Metrics: Population Count
006 3-Digit ZCTA	00631 5-Digit ZCTA	00 None	856
		1 Person	327
		02 People	388
		03 People	320
		04 People	173
		05 People	74
		06 People	32
		7 or more People	18
		Total	2,188
	Total		2,188
010 3-Digit ZCTA	01032 5-Digit ZCTA	00 None	138
		1 Person	24
		02 People	35
		03 People	15
		Total	212
	Total		212
Total			2,400

For confidentiality reasons, data may have been withheld from this tabulation. For a list of geographic areas which did not pass confidentiality, click here.

Source: U.S. Census Bureau, Census 2000 Hundred Percent Detail File
Data users who create their own tabulations using data from the Census 2000 Hundred Percent Detail File should cite the Census Bureau as the source of the original data only.

Fig. 3. Example of a Tier 3 data product

54 years, 55 to 59 years, 60 and 61 years, 62 to 64 years, 65 and 66 years, 67 to 69 years, 70 to 74 years, 75 to 79 years, 80 to 84 years, 85 years and over.)

All continuous variables, such as household/family income, individual income types, cost of electricity, gas, water, and fuel, property taxes, mortgage payments and gross rent, are top-coded. Top-coding is used to mask outlying values in the tails of distributions of continuous variables. For example, we would never show an income value of one million dollars. Instead we may set a top-code for income of $150,000 and make the highest category shown for income $150,000 or more. There are no hard and fast rules for determining which cutoffs to use in top-coding. Decisions should be based on examination of the structure of the distribution, in combination with other key variables like race, gender, etc. For example one may top-code the top 5% of the nonzero values.

Recoded variables are added to the files used by both American FactFinder and the Advanced Query System. In the Advanced Query System re-coded variables are accessed according to the query and who is making the query. Internal Census Bureau users may use all of the variables and geographic areas, as required by their work. External users are diverted to recoded variables and geographic areas of a certain size, depending on confidentiality issues such as:

- the population universe, geography and variables requested
- the rules and population thresholds the Census Bureau requires for the tabulation as a whole and the cells in the tabulation to meet confidentiality requirements.

3.2. Data swapping

A swapping technique was used for the 1990 Census of Population and Housing. A modified version of this technique was used for the Census 2000 one hundred percent (short form) data and independently for the Census 2000 sample (long form) data. The technique consists of swapping pairs of household records selected as having the greatest disclosure risk of being reidentified. In particular, records that are unique with respect to a set of variables are marked for swapping.

The variables that make a record unique are referred to as key variables. A record will be selected for swapping with a probability inversely proportional to block size. Records of households with unique race categories in the block will have an increased probability of being swapped. The swapped records match on a set of demographic characteristics but are in different census blocks for the hundred percent (short form) data and in different block groups for the sample (long form) data. The list of variables used to find the unique records and the list of variables used to find partnering households are both confidential. The majority of the variables were swapped, but a few that were tied to the geography and the housing unit were held fixed. For example, travel time to work and place of work for a household might not make sense if swapped with a household geographically far away.

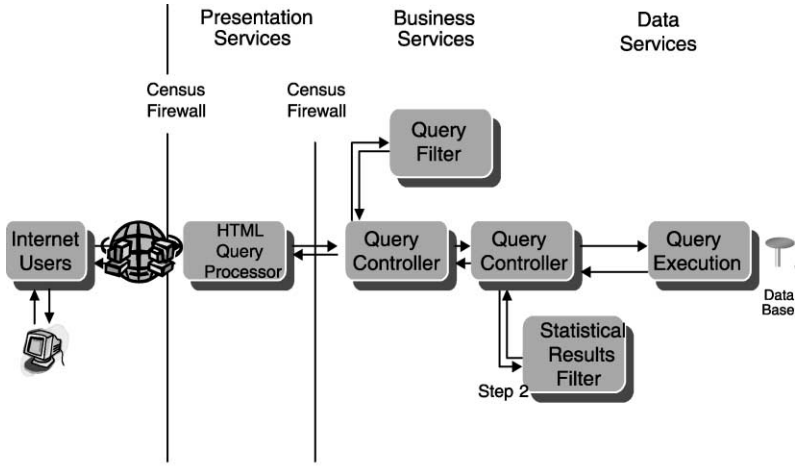
Both AFF and the Advanced Query System use the swapped data files as input. All tables publicly released (on paper, on tape, through AFF, etc.) to anyone outside the U.S. Census Bureau are generated from the swapped data files.

3.3. Advanced Query technology

Through the Advanced Query (AQ) system, the user can obtain very specific data for very specific geographic areas. Therefore, the AQ system is configured to permit web-service requests which originate on only the AQ external server, and which terminate on the AQ internal server. Communication from any other external machine to the internal server is blocked by two firewalls.

Any request for Tier 3 data from an external user is routed through the first firewall to the first server that contains the Query Processor, as shown in the illustration. By definition, there is no confidential data sent along this path. The external server receives requests for tabulations, which require confidential data, and retransmits them over the path through the second firewall to the second server that contains the Query Controller. Again, no confidential information is sent along this path. No traffic from the Internet can talk to the Internal Server because the firewall permits communication only between the two AQ servers. Completed tabulations are routed back through the Query Controller, which calculates the statistical results filter and sends the tabulations back to the user only if they pass the filters. If a table does not pass the filters, the user simply does not receive it. There is no use of cell suppression. The entire table is suppressed, and a message is sent to the user saying that the table is suppressed for confidentiality reasons. The user can then request a table with less detail than the original.

We discuss next the disclosure limitation rules designed for the AQ system. The rules are implemented as a pair of filters. There are two filters: the Query filter and the Statistical Results filter.



The Advanced Query System

3.3.1. Query filter

The purpose of the query filter is to detect those queries that will not pass disclosure limitation before they are submitted for execution. This saves the Tier 3 system resources and saves time for external users by telling them relatively quickly whether or not their query has a chance to pass disclosure limitation rules.

- Cross-tabulations must be created from a U.S. Census Bureau's predefined list of geographic areas and nongeographic variables.
- The smallest allowed geographic areas are block-groups for one hundred percent (short form) data and tracts for sample (long form) data.
- The query's geographic variable must meet a minimum population threshold.
- The system determines if the query requests small areas, block-groups or user-defined geography with a population size that is less than average tract size (population of less than 4,300), medium areas (population of 4,300 to 99,999) or large areas (population of 100,000 or more.)
- According to the population size of the area or areas requested, the system permits the use of appropriate combinations of short, medium or long lists of predefined categories of race, Hispanic origin, group quarters and other sample variables in the cross-tabulation. Only top-coded variables may be accessed.
- The maximum number of variables used to create an overall table is three, excluding the geographic variable and the universe.
- Depending on the cross-tabulation variables, the user must select from a list of derived measures possible (means, medians, . . .) to be reported within a cell.
- Derived measures are provided only if the corresponding counts are provided.

If a query passes all the disclosure limitation rules for the query filter, it is passed through the query controllers to the microdata files for computation. The microdata files contain all of the predefined categories for race, Hispanic origin, group-quarters, and modified sample data variables.

3.3.2. The statistical results filter

The results filter provides a final check on the values in the cells of the resulting table. The filter is designed to prevent the passage of tables that are very sparse, i.e., with more than a majority of the cells having a 0 or a 1. It serves as a reminder here that many of the predefined tables still contain cell values of 1, and for those we rely on the data swapping procedure to protect respondents from being reidentified.

- When geographic subtables are requested in the query, the disclosure limitation rules are checked separately for each geographic area – corresponding to the geographic subtable – as if the user had submitted separate queries, one for each geographic area.
- The median cell size in a requested table cannot be less than a parameter set by the Bureau.
- The mean cell size in a requested table cannot be less than a parameter set by the Bureau.
- The ratio of the number of cells with an unweighted count equal to one (note that all sample data tables are weighted) to the total number of cells in a requested table cannot be more than a parameter set by the Bureau.
- When a user defines recodes for a query, the AQ system first computes the results as if no recodes had been requested, and checks these results against the disclosure limitation rules. Following that, if the results satisfy the rules, they are aggregated to reflect the recodes requested.

The parameters used in the results filter are confidential. They were chosen to allow approximately the same amount of detail available in Tier 2 data products.

4. Testing and Current Status

The U.S. Census Bureau hired statistical disclosure limitation experts at Carnegie Mellon University to test the system (Duncan 2000). They found the confidentiality protections currently in place to be "thoughtful and comprehensive, and substantially consistent with best practice." The Bureau also hired a contractor to conduct a Beta Test for utility with a large set of users of decennial census tabular data products (Schneider 2002). She reported: "Of critical importance is the fact that the AQ system provided useful results. About 90 percent of the tabulations attempted either fully or partially met the intended objective. Testers expressed some frustration with the confidentiality filters but the filters did not seem to be a significant impediment to obtaining usable results."

In testing the system, we found that the requirement for a minimum mean cell size in the results filter was superfluous. That is, whenever this requirement was not met, one (or more) of the other requirements was also not met. Thus this requirement will be removed from the system.

Currently only the Beta Testers, the State Data Centers, and the Census Information Centers have access to the Advanced Query System. These users have extensive experience using U.S. decennial census data. They are familiar with the geographic summary levels and demographic characteristics used in the U.S. Census Summary Files produced for U.S. Census 2000. Before using the system, users are encouraged to first review the extensive Census 2000 summary tabulations provided in American FactFinder (factfinder.census.gov) and on CD-ROM. If the tabulations do not satisfy users' needs, then the use of the AQ system is appropriate.

There are concerns that if the system is opened up to the general public, the volume of queries might overload and shut down the system. The Bureau may expand the list of users, but no formal plan has been developed at this time. Also, currently, the system is free.

There is continuous use of the system (8.00 a.m.–8.00 p.m., Monday–Friday). Each day the list of external users is captured, along with the counts of tabulations they submitted and those that were not returned due to the confidentiality filters. Owing to limited resources, system managers have not yet been able to analyze these logs of current uses, but they plan to in the near future. The overhead is mainly following up on the user feedback and resetting forgotten passwords. Also, system managers are monitoring the way queries are being formulated by the SQL engine and adjusting the system to run better.

Current users seem very pleased with the system. When it was recently temporarily shut down for modification, we received numerous complaints and requests to get it running again as soon as possible.

The Advanced Query System has proven to be a very valuable tool for data users, and one that employs good disclosure limitation techniques to protect the confidentiality of U.S. Census Bureau data.

5. References

Blakemore, M. (2001). The Potential and Perils of Remote Access. 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., 315–340.

Duncan, G. (2000). Final Report on the American FactFinder Disclosure Audit Project for the U.S. Census Bureau. Prepared under contract to the U.S. Census Bureau.

Schneider, P.J. (2002). American FactFinder Advanced Query System – Assessment Report on Stage Two (Sample File) Beta Testing. Prepared under contract to the U.S. Census Bureau.

Zayatz, L. (2003). Disclosure Limitation for Census 2000 Tabular Data. Presented at the Joint European Commission for Europe and EUROSTAT Work Session on Statistical Data Confidentiality. http://www.unece.org/stats/documents/2003/04/confidentiality/wp.15.e.pdf

Received June 2002
Revised June 2003