Confidentiality Issues at the United States Bureau of the Census

Lawrence H. Cox, Sarah-Kathryn McDonald, and Dawn Nelson

Abstract: The U.S. Census Bureau undertook an intensive assessment of its confidentiality policies and practices during 1984. These studies covered disclosure avoidance for the 1990 census and other data releases, microdata for public use, interagency data sharing, computer security, and questions of public perception and trust. These studies pointed out the need for an integrated program of research and development work to contribute to key decisions on important policy issues. This paper presents the major ideas that were raised in these studies and merit further attention as opportunities for research, and highlights research in progress.

Key words: Disclosure avoidance; data sharing; public perception; confidentiality.

1. Introduction

The challenge posed by the dual concerns for privacy and the need to know is met by not collecting unnecessary information and by maintaining the confidentiality of information which is collected from individual respondents (U.S. Department of Commerce (1978)). The U.S. Census Bureau must balance the implementation of confidentiality requirements against the responsibility of being America’s major data supplier. The purpose of this paper is to provide information on the U.S. Census Bureau’s current confidentiality practices, its proposed plans, suggestions for further research, and on work in progress.

1.1. The U.S. Census Bureau as Data Provider

The Census Bureau occupies a preeminent position as a collector and purveyor of statistical information. The agency was created in 1902 to provide a permanent and professional staff to conduct the only statistical program specified in the Constitution – the decennial census of population – as well as economic, agricultural, housing and governmental censuses. In the 1940s, the agency played an important role in developing practical sample survey theory and methods. As a result, the Census Bureau’s program expanded to include
surveys in addition to the censuses, thereby greatly increasing the amount of information it collects. Today, it performs a major portion of its survey work on a reimbursable basis for other U.S. Federal agencies.

Thus, the U.S. Census Bureau is a major source of statistical information that is not available elsewhere and that has a wide variety of uses in both the public and private sectors. Historically, the primary mission of the Census Bureau was to provide data for the apportionment of political representation; however, today the same data are also used to meet a multitude of other important needs. Decennial census data are used to distribute U.S. Federal and state funds and for redistricting purposes. In addition, decennial census data users may include a market analyst for a firm who is selecting new store sites, a social science student conducting research at a university, or a government planner compiling data to determine which communities are eligible for a housing development grant.

Although the decennial census program is probably the best known, there are other data collection programs at the U.S. Census Bureau. Among the reimbursable surveys, a key program is the Current Population Survey, which collects labor force data. These data are used by the Bureau of Labor Statistics to produce the official U.S. monthly unemployment statistics. The Census Bureau also collects survey data on subjects such as income, education, health, crime, and expenditures for several other agencies. Other major programs include the economic censuses and current business surveys. These programs provide useful indicators of the Nation’s economy (e.g., housing starts and measures of U.S. Federal assistance to state and local governments) in addition to information useful to the individual businessperson.

The demand for detailed information is continually increasing, especially for small areas such as blocks and postal codes. As society grows more complex, the U.S. Federal government increasingly relies on statistics to allocate funds (such as welfare benefits); to regulate industry; to plan and evaluate programs; and to enact laws. Consequently, a tremendous amount of data is collected, some of it highly sensitive or capable of causing harm to the respondent if made public in its original form. One way of reducing the amount of data requested from respondents is to complement survey data with data from administrative records. This practice, however, is often perceived as a threat to confidentiality.

1.2. The U.S. Census Bureau’s Responsibility for Data Confidentiality

The U.S. Census Bureau collects the overwhelming majority of its data under Title 13 of the U.S. Code, the enabling legislation for the Census Bureau and its programs. According to Title 13, Sections 8 and 9, the collected data can be used for statistical purposes only and cannot be published in any manner that would lead to the identification of any particular establishment or individual. Methods of preventing the release of confidential data is a subject of great concern to the Census Bureau.

Prior to the 1960s, the U.S. Census Bureau released data from censuses and surveys that were for the most part in the form of tabular summary data. In the 1960s, it began releasing census and survey microdata tapes containing individual machine-readable records from which unique identifiers such as name and address were removed. Before releasing data products derived from confidential data, the Census Bureau must subject them to procedures that provide reasonable assurance that individual respondent data cannot be determined. However, as discussed in Section 2, these measures do not completely eliminate the risk of disclosure. Measures designed to completely eliminate this risk would seriously
curtail the number of and usefulness of statistical releases.

Three different groups of U.S. Census Bureau employees recently reexamined our confidentiality requirements and their effect on our role as the nation’s data provider. One group looked at current Bureau-wide policies and practices and recommended actions to correct inconsistencies and inadequacies. Another group concentrated on disclosure avoidance methods and confidentiality and data security issues related to preparations for the 1990 Decennial Census. The third group took a broad view of the Census Bureau’s long-term confidentiality issues and options within a strategic planning framework. This paper deals with the primary concerns of these three groups. The major outcome of these studies was the creation of a Confidentiality Staff to pursue research and development of confidentiality protection methods.

1.3. Conceptual Framework of the Paper

This paper discusses three types of confidentiality problems: 1) direct threats, 2) perceived threats, and 3) indirect threats. Direct threats encompass those situations in which confidential information is revealed directly to someone outside the U.S. Census Bureau. This can occur in two ways: 1) as a result of unauthorized actions, or 2) inadvertently through release of data, in either print or tape format.

Throughout its history, the U.S. Census Bureau has implemented many measures to counteract direct threats. Employees are required to take an oath, which they must renew periodically, swearing that they will not reveal confidential information to others. Interviewers are instructed on ways to avoid problems in certain interview situations, such as the presence of others during an interview or the use of a party-line telephone. Access to individual questionnaires and other confidential information is strictly limited at field and headquarters offices. This paper will discuss in detail two other measures of preventing direct threats. Section 2 presents techniques currently used to avoid disclosure in data products and proposes other techniques for consideration. Section 3 presents options being explored by the U.S. Census Bureau to prevent unauthorized access to computerized confidential information. These sections of the paper point out that additional research is still necessary to assess the effectiveness of the confidentiality measures taken.

The second type of confidentiality problem is perceived threats. An increased public awareness of the power of computers in conjunction with a decreased trust in governmental institutions has exacerbated the U.S. Census Bureau’s problem of maintaining the public’s trust. In this climate of suspicion, people may be threatened by plans under consideration for the 1990 Decennial Census or other Census Bureau programs that involve more widespread use of computers and increased use of administrative records. Although appropriate policies and practices will be implemented to continue to ensure a respondent’s confidentiality, the Census Bureau cannot expect all members of the public to know, understand or believe the ways these measures protect them. Perceived threats existed in the past, but not to the extent that resources were required to deal with them. To ensure public cooperation in the future, however, the Census Bureau may have to expend resources to address perceived threats in addition to known actual threats. Section 4 highlights some of the research opportunities that exist in this area.

The last section of this paper briefly discusses the most difficult type of problem for the U.S. Census Bureau to control, i.e., indirect threats. Indirect threats result from others who use or manipulate census data in ways that appear to compromise confidentiality. Summary data, for example, can be used in ways that affect an entire group or class of
people as opposed to a single individual. An example of indirect threats is the use of summary data to deliver specialized commercial mail to certain postal codes. In such cases, it may appear that a breach of confidentiality has occurred. This problem is difficult to address since the Census Bureau cannot limit the purposes for which its data are used once they are released in a form that does not identify an individual. Section 5 directs attention to this little known or understood problem in hopes that it will prompt further research.

This paper is designed to stimulate the interest of others in the U.S. Census Bureau's research plans and needs in the area of confidentiality. By sharing this information, the Census Bureau hopes to learn more about current research outside the Bureau. Examples include the work of Paas (1985) that explores the difficulties of disclosure-avoidance for microdata and that of Duncan and Lambert (1986) that considers distributional threats to confidentiality. The Census Bureau hopes this paper will stimulate discussion of these issues and encourage further research.

2. Confidentiality in the Release of Data Products

As previously stated, direct threats may occur inadvertently through the release of data. Therefore, before releasing data products derived from confidential data, the U.S. Census Bureau must take steps to thwart identification of individual respondent data, in conformance with Title 13 and other confidentiality guidelines. This process is called disclosure avoidance. In the strictest sense, 100-percent "avoidance" of disclosure is impossible by any reasonable standard. What disclosure avoidance methods really do is abbreviate or otherwise distort the released data sufficiently to reduce the risk of identifying individual respondents and divulging their confidential data. The challenge posed by the confidentiality problem is to provide protection to respondents without damaging the completeness, accuracy and usefulness of the data product. This section deals with modern methods for disclosure avoidance, areas in which additional methodology is needed, and candidate methods for these deficient areas.

The U.S. Census Bureau releases three types of data products that are subjected to disclosure avoidance measures: sets of frequency count tables, sets of tables of aggregate magnitude data, and microdata files. Each poses a unique set of disclosure avoidance problems. Statistical tables are usually released as large, interrelated sets of tables, so that disclosure avoidance must be performed both within and between individual tables. Moreover, the tabulation structure often is such that well-defined groups of two-way tables combine to form three or even higher dimensional tables. The objective of disclosure avoidance methods for tabular data is to thwart disclosure of respondent data through analysis of the published cells and, in particular, through manipulation of linear relationships between them.

In the past, the U.S. Census Bureau performed disclosure avoidance in tabulations either by presenting data in broad or grouped ("collapsed") categories, withholding some data outright, or deleting ("suppressing") troublesome cells from publication. Before reliable methods were developed, disclosure avoidance was an ad hoc process, which sometimes was incomplete or inconsistent, paying unsystematic attention to the means by which it could be undone. Recently, more sophisticated methodology based upon established theoretical principles has been developed or postulated. Some of these methods have been implemented at the Census Bureau; research into others has begun.

Microdata records contain entirely disaggregated information for individual
respondents. Prior to release of a microdata file, the Census Bureau removes ("strips") all directly identifying information from all microrecords and curtails the geographic detail presented on them. Still, identification of individual respondents remains a possibility, and, typically, additional disclosure avoidance techniques must be applied.

The remainder of this section is devoted to examining in detail the threats to confidentiality posed by releasing each of these data types. This examination has helped to establish a preliminary research agenda for the Confidentiality Staff, created in 1985, on the recommendation of the study groups.

2.1. Disclosure Avoidance in Frequency Count Tabulations

Disclosure occurs in frequency counts when small counts are published or may be inferred and such values are used to identify individual respondents and divulge their characteristics. For example, if it is well-known that in a particular county there is precisely one female statistician, then release of frequency data for this county consisting of income category cross-classified by sex and occupation would constitute disclosure of this person's income category. Consequently, disclosure in frequency counts is defined by a threshold rule. Publishing the frequency count (the number of respondents or the value) of a specific statistical cell, such as the number of female statisticians in Fairfax County, Virginia, represents disclosure if the count is less than a predefined threshold value. Under a threshold disclosure rule, a cell or cell combination \( X \) is a disclosure cell if

\[
\# (X) < n
\]

where \( n \) is the threshold value and \( \# (X) \) is the frequency count for \( X \). The U.S. Census Bureau has used the threshold values \( n = 5 \) households and \( n = 15 \) persons in its decennial census publications.

Although single zero counts do not always represent disclosure, sets of zero counts can. For example, if all but one cell from an additive set of cells for an income distribution are zero, then all respondents must be in the same income category. If this category is narrowly defined, then detailed income information has been disclosed for each respondent. Historically, however, the U.S. Census Bureau has not treated zero cells as representing disclosure. In the context of examining agency disclosure rules, this practice will be reviewed. Cell suppression methods suppress the disclosure cells from publication, together with additional nondisclosure cells (called the complementary suppressions) chosen to thwart reconstruction or narrow estimation of the disclosure cells by means of linear relationships between the cell values. In the 1980 and previous decennial censuses, the Census Bureau employed ad hoc complementary cell suppression procedures to avoid disclosure in its frequency count tables. Not based upon established statistical methodology, these procedures were not 100 percent reliable. Also, based upon policy and program considerations, their use was restricted: only marginal totals in the frequency distributions were examined for disclosure potential; some tabulations were not subjected to any disclosure avoidance; and complementary suppression was not applied in all cases (such as across geographic areas).

These policies and practices were developed on the basis of sound data use considerations, and were necessitated by the poor complementary suppression methods available. These methods and practices also resulted in inconsistencies. There was oversuppression in some tables (due to the use of imperfect ad hoc suppression strategies) and undersuppression in other tables (due to subjective determinations and technical limitations of the ad
hoc strategies). Fortunately, the state-of-the-art has advanced since that time (Cox (1980, 1983), Gusfield (1984)). Four options have been identified for the choice of a statistically defensible disclosure avoidance methodology for the 1990 census frequency count tables: cell suppression (based upon established theory, not ad hoc procedures), random data perturbation, random rounding and controlled rounding. We restrict attention to single two-way tables of frequency counts, as the problem of disclosure between tables or in higher dimensional tables is more complex and beyond the scope of this paper. We illustrate the four candidate methods by means of the following two-way frequency count table.

\[
\begin{array}{|c|c|c|c|c|}
\hline
D_1 & D_2 & D_3 & 15 & 20 \\
15 & D_4 & D_5 & 20 & 55 \\
D_6 & 10 & 10 & D_7 & 25 \\
D_8 & 6 & 15 & D_9 & 35 \\
\hline
20 & 30 & 35 & 50 & 135 \\
\hline
\end{array}
\]

(2)

Each row and column of (2) provides a linear equation in the suppressed entries (the "D's"). By subtracting the sum of the equations provided by the second and third columns from the sum of those provided by the first and second rows, we are able to discover \(D_1\), the value of the cell in the upper left-hand corner of (2), precisely. A linear programming analysis of (2) would enable us to find all unprotected disclosures systematically. To ensure that adequate disclosure protection is provided while suppressing the least amount of data possible, the linear analysis or an equivalent should be an integral part of the cell suppression strategy.

Table (2) fails in part because not all linear combinations of the suppressed cells contain two or more variables. Gusfield (1984) has solved the problem of extending a suppression pattern to satisfy this condition by posing it as a graph augmentation problem. Research is needed to extend these results to the full context of the practical cell suppression problem as described in Cox (1980, 1983), and to develop similar methods for sets of two-way tables and higher dimensional tables. For table (1), the cell suppression pattern (3) is optimal in two senses: it suppresses the fewest cells possible and also minimizes the total value of the suppressed cells. However, in general both of these objectives cannot be satisfied simultaneously.

\[
\begin{array}{|c|c|c|c|c|}
\hline
D & D & D & 15 & 20 \\
15 & D & 10 & 10 & 55 \\
D & 10 & D & D & 25 \\
D & D & D & 15 & 35 \\
\hline
20 & 30 & 35 & 50 & 135 \\
\hline
\end{array}
\]

(3)

For a threshold value \(n = 5\), the table (1) contains six disclosure cells (note: we assume here that zero cells do constitute disclosure). These six cells are in parentheses.

Under cell suppression, the six disclosure cells would be suppressed from publication, together with additional, nondisclosure cells to prevent identification of the disclosure cells through manipulation of linear relationships between the cells. An ad hoc cell suppression methodology typically will miss the mark: either it will fail to adequately mask the values of the disclosure cells and/or it will suppress too much information. For example, it may appear that release of the table (2) provides the requisite disclosure protection, but it does not.
Data perturbation methods distort the values of all (or most) cells in a frequency count table slightly to prevent disclosure of confidential data for any single cell. Random perturbation works as follows. A fixed set $V$ of perturbation values $\{v_i\}_{i=1}^m$ and an associated set $P$ of probabilities of selection $\{p_i\}_{i=1}^m$, $\sum_{i=1}^m p_i = 1$, are chosen for which the expected perturbation is zero, i.e., $E(\nu) = \sum_{i=1}^m p_i \nu_i = 0$. Typically, $V$ is the set of integers from $-t$ to $t$. For each cell $X$, a random choice $\nu$ is made from $V$ and the value $\max(\#(X) + \nu, 0)$ is published in place of $\#(X)$. The Office of Population Censuses and Surveys of Great Britain uses random perturbation with $t=1$ and $P = \{1/4, 1/2, 1/4\}$ (Newman (1975)). Expected values are maintained for some but not all aggregates due to the upward bias introduced by imputing zeros for counts with negative perturbed values. In general, random perturbation destroys additivity to totals.\(^2\)

To illustrate the use of random perturbation on table (1) we set $t=3$ and $P = \{1/10, 1/10, 1/10, 4/10, 1/10, 1/10, 1/10\}$. To preserve additivity to totals we pair internal cells along the rows, pair and perturb marginal cells similarly, and leave the grand total unpaired. The result of this procedure applied to table (1) is table (4).

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5</td>
<td>0</td>
<td>15</td>
<td>19</td>
</tr>
<tr>
<td>14</td>
<td>11</td>
<td>11</td>
<td>19</td>
<td>56</td>
</tr>
<tr>
<td>0</td>
<td>13</td>
<td>10</td>
<td>3</td>
<td>26</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>13</td>
<td>14</td>
<td>33</td>
</tr>
<tr>
<td>21</td>
<td>29</td>
<td>35</td>
<td>50</td>
<td>133</td>
</tr>
</tbody>
</table>

Table (4) is not additive. Also, there is the general problem that, whenever the published value equals 0 or 1, identification of the cell as a disclosure cell can be made with certainty. This can be overcome by choosing $t > n$, but at the risk of introducing unacceptably large perturbation values which will affect data accuracy considerably, particularly for small counts.

Rounding can be used to mask the true cell values. However, conventional rounding is not a viable method for disclosure control. Under conventional rounding, with a fixed base $B = 5$, we round the frequency counts down to the nearest multiple of 5 wherever the remainder modulo 5 of the number is equal to 0, 1 or 2, and up to the nearest multiple of 5 whenever the remainder modulo 5 of the number is 3 or 4. Although conventional rounding minimizes standard measures of deviation between the original and rounded arrays, it has two distinct shortcomings. First, it does not maintain additivity. Second, because the conventional rounding of a table is unique, it sometimes can be undone, i.e., from certain conventionally rounded arrays, such as (5) below, the original table can be reconstructed with certainty.

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>10</td>
</tr>
</tbody>
</table>

To avoid such situations, more than one choice for each rounded value must exist. The obvious choice for these values are the 2 integer multiples of the base $B$ which are adjacent to the unrounded value.

Random rounding can be viewed conceptually as a variant of random perturbation. The rounding is done to a fixed base $B$; for con-
creteness, we set $B=5$. Random rounding works as follows. For each cell $X$ in the table, the remainder $r$ of $\#(X)$ modulo 5 is computed (i.e., write $\#(X) = 5q + r$ for integers $q$, $r$ with $0 \leq r < 5$). In place of $\#(X)$, either $5q$ or $5(q+1)$ is published, i.e., $\#(X)$ is rounded either down or up to an adjacent integer multiple of 5. This rounding, however, is done probabilistically. We round down (up) with probability $(1-r/B)$ (probability $r/B$). Thus, in random rounding each cell frequency is rounded to an adjacent integer multiple of 5. The choice of probabilities ensures that the rounding is unbiased, i.e., the expected value of the rounded count equals the original count. In the general case, by choosing a fixed base equal to the predefined threshold value ($B=n$), all published values are multiples of the threshold, and hence precise determination of each frequency count is rarely possible. Still, the determination $\#(X)<n$ can be made with certainty wherever the rounded value equals 0, but this can be avoided by choosing a base greater than the threshold value ($B\geq n+1$).

In one-way tables, random rounding can be controlled so that the sum of the rounded entries equals the rounded value of the sum of the unrounded entries (Fellegi (1975)). Unfortunately, this random rounding procedure cannot be extended to two-way tables in a manner that guarantees the controlled property, so that, in general random rounding applied to a two-way table produces an array of rounded values but not an additive two-way table. Also, beyond unbiasedness, random rounding offers no control over the deviation between the rounded and unrounded arrays, as measured in any of several standard ways (e.g., sum of (squared) absolute differences between original and rounded entries).

Random rounding with a fixed base $B=n=5$ applied to the entries of table (1) yields table (6).

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>15</td>
<td>10</td>
<td>10</td>
<td>20</td>
<td>55</td>
</tr>
<tr>
<td>0</td>
<td>10</td>
<td>10</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>0</td>
<td>10</td>
<td>15</td>
<td>15</td>
<td>35</td>
</tr>
<tr>
<td>20</td>
<td>30</td>
<td>35</td>
<td>50</td>
<td>135</td>
</tr>
</tbody>
</table>

Application of controlled random rounding along, say, the rows of table (1) would improve additivity somewhat but not entirely.

*Controlled rounding* was developed to overcome the shortcomings of conventional and random rounding and to combine their desirable features. Like random rounding, controlled rounding replaces the original two-way table by an array whose entries are rounded values adjacent to the corresponding original values. However, under controlled rounding, the rounded array is guaranteed to be additive in both dimensions and, subject to this requirement, can be performed optimally, i.e., can be chosen to minimize any of a class of standard measures of deviation between the original and rounded tables (Cox and Ernst (1982)).

Controlled rounding can be achieved by means of a specialized linear programming model, the transportation model, for which standard, efficient computer software is available commercially (Cox and Ernst (1982)). As with random rounding, we must require that the base is greater than the threshold value; otherwise, frequency counts in the forbidden range $0-4$ can be inferred from rounded values which equal 0. Moreover, unbiased controlled rounding is possible. This is done by a well-defined procedure that constructs a set $R$ of controlled roundings, together with associated probabilities of selection $P$, for which the expected value property holds for all entries in the table, and randomly selecting one controlled rounding from $R$ according to the probability distribution $P$ (Causey, Cox, and Ernst (1985)). Unbiased
Controlled rounding can also be achieved by a much simpler algorithm requiring only one rounding step (Cox (1985)). Controlled rounding applied to table (1) yields table (7).

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5</td>
<td>0</td>
<td>15</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>10</td>
<td>10</td>
<td>20</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>10</td>
<td>10</td>
<td>5</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>15</td>
<td>10</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>30</td>
<td>35</td>
<td>50</td>
<td>135</td>
<td></td>
</tr>
</tbody>
</table>

(7)

Each of the four methods for disclosure avoidance in frequency count tables above is viable for single two-way tables. The next level of complexity to be considered concerns sets of (two-way) tables: Can methods be applied consistently between tables (i.e., so that rounded counts for equivalent cells are equal)? The objectives of additivity within tables and consistency between tables conflict. Random perturbation and random rounding can be applied consistently, but at the loss of additivity. Examples show that controlled rounding cannot always be applied consistently between sets of tables. With some effort, cell suppression can be applied consistently between sets of tables but creates problems for data use; problems which increase with the complexity and number of levels of aggregation and the vigor with which the suppression is applied.

Among the three nonsuppression methods currently implemented, only controlled rounding meets all of the following objectives: additivity, unbiasedness and, subject to the first two objectives, reduction of data distortion. The properties of all four methods as currently implemented are summarized below.

There are several research directions for disclosure avoidance in frequency counts. For the U.S. Census Bureau, the broad objective of this research is to identify and develop a single, preferred method of disclosure avoidance for frequency counts, or to demonstrate that narrowing the choices to one method is unwise. The Census Bureau plans to fully explore, develop, and test the four alternative methods, as well as meaningful combinations of these methods. To conduct a full exploration for cell suppression, it needs to test the efficiency of new methods such as those of Gusfield (1984) and to extend them to the full range of theoretical and practical disclosure avoidance problems for frequency counts. Random perturbation and random rounding, being fairly well-established, require little or no theoretical work but rather refinements and efficiencies obtained

<table>
<thead>
<tr>
<th></th>
<th>data perturbation</th>
<th>random rounding</th>
<th>controlled rounding</th>
<th>suppression</th>
</tr>
</thead>
<tbody>
<tr>
<td>additivity</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>unbiasedness</td>
<td>biased for small counts</td>
<td>yes</td>
<td>yes</td>
<td>n/a</td>
</tr>
<tr>
<td>reduction in data distortion</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>consistency between tables</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
</tbody>
</table>
empirically. Controlled rounding must be implemented efficiently and the extent to which it can be applied to higher dimensional tables and consistently between tables must be investigated theoretically. During 1986, the alternative methods will be studied, improved, and implemented for single tables. In 1987, generalizations to sets of tables and to higher dimensional tables will be pursued.

2.2. Disclosure Avoidance in Tabulations of Aggregate Magnitude Data

Aggregate magnitude data, such as sales, pay-roll and expenditures data for retail stores, is also presented in statistical tables. However, disclosure in aggregate magnitude data is qualitatively different from that for frequency count data. For magnitude data, it is not the presence of one or a small group of respondents in the cell that must be masked, but rather their contribution to the cell value. For example, publicly available information such as the Yellow Pages would allow the curious to know which retail stores contribute to a particular cell defined, say, at the county level in a statistical table drawn from the economic censuses. Disclosure would result if the curious person or firm could obtain a narrow estimate of the value of sales for one or more of these stores. The offending party could be one or a coalition of these firms who, after subtracting its total contribution from the published cell value, seeks to narrowly estimate the contribution of a competitor.

In its economic statistics, the U.S. Census Bureau employs a cell concentration rule to define disclosure: if \( n \) or fewer respondents contribute more than \( k \)-percent to the cell value, then the cell is considered to be a disclosure cell. The values of \( n \) and \( k \) are treated as Census-confidential information as a further safeguard to privacy (\( n \) here bears no relationship to the threshold values \( n \) discussed previously). Whereas there are several viable disclosure avoidance methods for frequency counts, no rounding or perturbation method is suitable for magnitude data, primarily because of the size of the values involved and the skewness of their distribution. For example, the rounding necessary to protect a respondent contributing \$800,000\) to a cell of value \$1,000,000\) would be so excessive as to destroy the usefulness of the data entirely. Other than collapsing definitional categories (which is unsystematic), there are only two viable disclosure avoidance methods for aggregate magnitude data: cell suppression and replacement of cell values by value ranges. As we shall see, the second of these may be viewed as a by-product of the first.

Cell suppression for aggregate magnitude data is more complex than for frequency counts, for two reasons. First, for frequency counts an inference which provides a lower bound on \( \#(X) \), the value of a cell \( X \), is often harmless, whereas, for aggregate magnitude data, a narrow lower estimate of the value \( \#(X) \) of a disclosure cell is no less damaging than a narrow upper estimate. Therefore, cell suppression methodology must be equally effective in thwarting both types of estimation in this case. This requires sophisticated cell suppression strategies (as discussed previously) which are integrated with efficient mechanisms to estimate the value of suppressed cells (a disclosure audit capability). Such mechanisms could be based upon specialized linear programming or network flow algorithms. Another complication in aggregate magnitude data not found with frequency counts is that it is possible for one respondent to contribute to several different cells and to combinations (aggregates) of these cells. For example, a department store chain may do business in several counties in a state or a manufacturing conglomerate may be active in several lines of business. Disclosure avoidance must be effective in protecting respondent data in cell combinations. This involves a subcell analysis which must be integrated with the cell suppression and disclosure audit strategies.
Cell suppression methods developed at the U.S. Census Bureau for the 1977 Economic Censuses and used since that time are based upon combinatorial algorithms, applied in conjunction with linear programming analysis and examination of critical cell combinations (Cox (1980)). These methods have worked well and have desirable theoretical properties, but are imperfect. The work of Gusfield (1984) offers one avenue of improvement which will be explored. Another is to represent the cell suppression problem within a mathematical programming framework. Because such models typically are non-linear (due to the objective to minimize the number of cells suppressed), they are not effective for large problems such as censuses. Alternative linear models or specialized mathematical programming models and algorithms that are efficient for large problems are needed.

Once a complete cell suppression has been performed, the suppressed values are typically replaced by a symbol, such as “D.” Through linear programming analysis, a user could reconstruct intervals in which these values must fall. The U.S. Census Bureau could publish these intervals instead of the “D’s” in an effort to make the data as useful as possible. If it were possible to release intervals for all or most of the cell values, then simpler, linear programming models for cell suppression could be employed. If this is done, however, generalized tools for manipulating interval numbers should be developed for the data user. Also, the case of a cell representing precisely one respondent must be carefully considered from the standpoint of perception of disclosure.

2.3. Disclosure Avoidance in Microdata

A microdata record contains precise, detailed information collected from an individual respondent. Statistical disclosure occurs if a respondent can be linked to his/her microdata record correctly and unambiguously, with confidential data thereby divulged. Consequently, name, address and other directly identifying information must be removed from the microrecord. Microdata records are usually drawn as a sample from a larger data file or are the result of a sample survey which affords the respondent some degree of inherent confidentiality protection. Changes that take place between data collection and release and other inaccuracies in the data can also function to protect a respondent’s identity. Nonetheless, it is usually also necessary to collapse or blur geographic detail on the file to thwart attempts at identifying individual respondents.

In the 1960s, the U.S. Bureau of the Census adopted a practice whereby standard geographic detail below a population threshold of 250,000 persons would not be shown on a microdata record. In 1981, this cutoff was reduced to 100,000 persons. In addition, prior to its release the microdata file is scrambled out of geographic sort order.

Disclosure control methods in microdata aim at making the task of identifying a respondent from his/her microdata record infeasible or impractical, recognizing that, in general, this task cannot be made impossible if useful microdata are to be released. The above methods usually provide adequate disclosure protection for the majority of microdata respondents. However, good measures of microdata disclosure risk are lacking. The work of Duncan and Lambert (1986) and Paas (1985) promise to lead to such measures. For example, little has been done to quantify the relationship between disclosure reduction and sampling rates or population thresholds. Also, microdata records typically contain accurate detailed information that is either well-known publicly or can be matched to (computerized) files available outside the U.S. Census Bureau. For these reasons, even after stripping direct identifiers and abbreviating geographic detail, one or more respondents may be uniquely identified from the
microdata, so that further steps must be considered. These are described below.

The principal techniques of disclosure avoidance for microdata are to collapse the data or to perturb it. Collapsing means to represent the data categorically, e.g., age represented in 5 year groups or income to the nearest $5,000. Conceptually, one can view the stripping of identifying information from each microdata record as an extreme form of data collapse. Topcoding is a specialized form of collapse where, for example, all income values in excess of $100,000 are coded into a single "$100,000+" category. (The U.S. Census Bureau currently employs a $100,000 income topcode which is adjusted for inflation periodically.) In principle, if not always in practice, the determination of class boundaries, topcodes, and similar categories is made on the basis of distributional properties of the population or sample data.

Data perturbation and rounding are also used, but less frequently. For example, in 1977 the U.S. Census Bureau agreed to create a microdata file for use and dissemination by the Department of Energy. This file was constructed by matching the 1976 Annual Housing Survey (AHS) for selected geographic areas to energy use and cost data obtained from local utility companies. Because the energy cost data were reported to the penny for a 12-month period, it was necessary to distort the accuracy of the energy usage cost data prior to release to guard against the possibility that the utility companies could uniquely identify individuals on the released file by the cost data. The techniques investigated in this case were random perturbation (studied for perturbations drawn from several distributions) and rounding to the nearest two dollars (Clark (1978)). As a result of this study, random noise from uniform distributions was added to the energy cost data for the middle 95–97 percent of the distribution; observations in the upper distributional tails were replaced by the tail means. In addition, energy consumption data were altered so as to correspond with the cost data.

There are two types of nonrandom perturbation that promise to be useful: controlled data rounding and microaggregation. Controlled rounding can be applied to microdata, as follows. Suppose that the microdata file or a subset is drawn as a random sample from the population of interest such as households in the states of the United States. Assume that each microrecord contains several income items and their total. To achieve disclosure avoidance, one might round the exact income figures, but only slightly which would not distort unnecessarily inferences and estimates that could be made from the data. In particular, it would be desirable to do so in an unbiased manner and to preserve the total value of each income item at the state level, as well as preserving additivity of rounded detail income items to the total within each microrecord.

By creating a two-way table for each state which contains one row for each microrecord from this state and one column for each income item and the total and applying unbiased controlled rounding to an appropriately chosen base \( B, (e.g., B = \$1,000) \), one may achieve these objectives (Causey, Cox and Ernst (1985)).

Microaggregation amounts to creating suitable small sets of microrecords and creating one or more synthetic microrecords whose data values are derived from the original microrecords. The file of synthetic microrecords is released. Only the synthetic values may be means of the original values. For example, under the assumption of serial correlation, we may group microrecords geographically and replace all values in the microrecords by the mean of the corresponding subgroup. Spruill (1983) and recent extensions of that work offer some promising possibilities for the use of microaggregation.

Paass (1985) examined reidentification risk in microdata resulting from modern matching methods and demonstrated the vulnerability
of the methods in current use discussed above. This work stimulates needed research into microdata disclosure. Further research includes: investigating the applicability of controlled rounding; exploring the effects of sample design or selection on preserving respondent confidentiality; topcoding and the extent to which the distribution of data in top-coded categories can be made available; further study of microaggregation techniques; and the problem of setting suitable population thresholds. A major, virtually unexplored area of research for all three data types is to quantify the effects of various disclosure avoidance methods upon data use. This latter area is more complex. For microdata release, if too many variables are analyzed at once, every record can be unique (in terms of variable values, ratios of variables to each other, etc.). Disclosure avoidance methods, applied vigorously, can easily mask or suppress information of no use to the person trying to find the disclosures, since such a person would know "more" already. A proper balance between disclosure avoidance and data use must be found. Paass (1985) discussed some of these issues.

Some of the new uses being explored include automated questionnaire check-in, data capture and editing; the use of hand held microcomputers for address list compilation and failed-edit follow up; the use of computer assisted telephone interviewing also for follow up; automated access to a name and address file; automated matching of administrative lists or other surveys to the census for coverage evaluation and coverage improvement operations; automated coding of write-in entries and on-line data scan and retrieval. Many of these operations were conducted manually in the 1980 census. They required large numbers of clerical personnel, long periods of time for completion and, therefore, were very expensive.

Automation of these clerical operations will allow the census to become more efficient and reduce costs. Automation can also increase accuracy by reducing human intervention and therefore, the possibility of introducing human error in the operations. It will also permit the detection and resolution of problems while the field offices are still open. Finally, it can lead to an earlier release of data products than occurred after the 1980 census.

In 1980 there were only three centralized sites containing all the hardware and software necessary to process the questionnaires from the over 400 field offices, which were responsible for manually checking-in and editing the questionnaires for completeness, compiling preliminary population and housing counts, followup assignments, and so forth. The activities conducted in these processing centers included clerical coding of written responses on the questionnaire, microfilming, data capture, and review and correction of questionnaires that failed computer editing.

Two basic decentralized automation scenarios were considered for 1990: 1) 400 to 500 local offices supported by multiple processing offices (3 to 50), or 2) 400 to 500 self-contained local data collection/processing offices. In the first scenario the processing
offices would contain most of the computer hardware required for automation and would receive and process the questionnaires. The 500 local collection offices would become field followup offices. Local collection offices might contain some computer terminals to record or receive information necessary for address updating, followup and administrative activities. The second scenario would entail a completely decentralized processing system. The local data collection/processing offices would contain the hardware for check-in and data capture, in addition to conducting field followup and administrative activities. Variations of each scenario are possible. For example, in the second scenario the sample data could be passed on to a regional processing office for data capture. Each of these scenarios could entail the use of an extensive communications network for the transmission of census data to and from numerous locations. Recently, the decision was reached to combine these scenarios for 1990.

The decentralized automation of census processes and the electronic storage and transmission of data present new and different direct threats because of the increased potential for penetration of security and abuse of the data. Also, management control of the operations and the flow of communications may be harder to maintain. The potential for unauthorized access by outsiders as well as insiders increases with the geographical distribution of the system, the amount of data available through remote access equipment, and the number of persons allowed access to the equipment and files. The interception of data communications, which would allow intruders to modify, destroy or obtain data from the system, must be considered a possibility. The need for an automated name and address file will require more attention to communications security. Automated access to names and addresses—even though just available temporarily until the census is completed—may have an effect on the public's perception of the confidentiality of the data. On the positive side, automation might help solve old security problems such as reducing the handling of questionnaires by many employees during the various clerical operations and possibly eliminating the need for enumerators to carry the questionnaires during followup operations.

Adequate physical security measures will be required for those sites selected for data collection and processing offices. These measures are needed to avoid the theft of the equipment, not only because of its monetary value but because of the potential loss of data. Such measures will also limit physical access to the computers, the stored census data, and other confidential census materials.

All of the previously mentioned difficulties are potential problems that must be seriously considered by the U.S. Census Bureau before final plans are made and implemented.

3.2. Solutions to Confidentiality Problems Associated with New Technologies

In view of the above mentioned potential security problems, before implementing its automation plans for 1990, the U.S. Census Bureau will have to establish systems that continue to guarantee the confidentiality of the data, as well as a clear and consistent policy for the transmission of confidential data through various levels of the system.

The private sector and other U.S. Federal agencies have been doing research in the areas of computer security enhancement, security vulnerabilities and risk analysis. For example, a study on security and vulnerability of Federal government information systems was initiated by the Congressional Office of Technology Assessment in September 1984 (Kirchner (1984)). As a result of these research efforts, several measures exist already that could be taken to protect the confidentiality of census data in this automa-
ted environment. In terms of the existing physical security measures, there are devices that may be used to avoid the theft of software and hardware. Other security measures that might be taken (some of which were implemented in 1980) involve preventing physical access as well as system access to the equipment and files. For instance, the U.S. Census Bureau's mainframe computers are not accessible by telephone lines from outside. The use of floppy disks and their drives could be avoided since it is very easy to steal or copy a disk. The computer storage media (i.e., tapes, disks) could be in a physically secured area accessible only to those employees responsible for them. Special badges could be worn by census employees with authorized access to ADP equipment and files. Access to the office sites could be carefully controlled. There could be a record of data access to hold employees accountable.

Technology already exists to enhance the level of the computer system security. System access can be prevented by the use of a software system using unique identification numbers and passwords in conjunction with some state-of-the-art hardware, such as devices to identify fingerprints, magnetic badges and so forth. Types of communications security include the use of dedicated lines and encryption devices, i.e., private lines dedicated exclusively for the U.S. Census Bureau and scrambling the data with a key encryption system before sending it over the communication lines. Public communications networks have some level of security built into the system so that it is extremely difficult to tap information from the network lines. Since the name and address file is separate from the file containing data items, confidentiality is, in fact, protected. A perception problem, however, may still remain. As an additional precautionary measure, names could be encrypted and matched with encrypted data. Finally, the names would be erased once all processing operations are completed.

The U.S. Census Bureau needs to conduct an in-depth analysis of the security vulnerabilities of the automation configurations being considered for 1990 and to develop the guidelines for implementation of the necessary security measures. Such a study would involve a risk analysis that would enable the U.S. Census Bureau to assess the likelihood of disclosure, the effect on Census Bureau operations if disclosure occurred and what management is willing to spend to prevent disclosure (Campbell (1984)). The security measures considered as a result of this analysis need to be evaluated in terms of real security needs versus perceived ones, cost and time effectiveness, ease of use and implementation, and so forth. The U.S. Census Bureau will test some of these security measures during the decennial census test cycle. The U.S. Census Bureau could greatly benefit by sharing expertise in this area with others already engaged in this research area.

All of these field and computer security measures are important also in terms of public perception, in light of the increased concerns displayed by the media and the public with computer crime. In addition to actually protecting the confidentiality of the data it collects, the U.S. Census Bureau needs to preserve the public perception that Census Bureau security is adequate. The area of computer technology and public perception is a fertile one for research. It is discussed in a broader context in the following section.

4. Public Perception and Willingness to Cooperate with Census and Survey Activities

As discussed in the two previous sections, direct threats are the easiest for the Census Bureau to understand and, through research, to counteract (e.g., through disclosure avoidance techniques; headquarters, field and computer security measures). Perceived threats are more difficult both to understand and to address.
Perceived threats are not necessarily grounded in fact. They may arise as a consequence of general attitudes, opinions or beliefs that are inappropriately applied to the U.S. Census Bureau. They may arise as a result of the dissemination of incomplete or inaccurate information about census taking procedures. They may also result from lack of information about the census and the Census Bureau’s commitment to maintaining the confidential nature of the individually identifiable data it collects. Depending on the circumstances, perceived threats may very closely match or may bear little resemblance to “actual” direct or indirect threats to confidentiality. Whatever their cause, and however accurately they mirror reality, perceived threats can have just as real an effect on the Census Bureau’s operations as direct and indirect threats.

Privacy and confidentiality are distinct concepts; one refers to the nature and the other to the status of information. Nevertheless, in the area of perceived threats, these two concepts become intertwined. The key concept underlying the notion of a right to informational privacy is control, specifically, the right to control access to or the dissemination of information. To the extent that an information collector disseminates information without the respondent’s knowledge or consent, a breach of confidentiality has occurred, and the respondent may well feel that an invasion of his/her privacy has taken place. In such a case, the information collector has violated the respondent’s trust, and effectively eliminated the respondent’s control of his/her individual information. Information which the respondent does not object to providing to one group or individual, he/she may object to providing to another; thus the wider dissemination of that information may simultaneously involve both a breach of confidentiality and an invasion of privacy.

4.1. Perception of the Need for Information

Demonstrating the legitimate need for information is not always an easy task. The U.S. Census Bureau may not be able to convey to everyone an appreciation of the utility and wisdom of the services it performs. Too often the questions asked by the Census Bureau may seem to be an invasion of privacy because some people do not recognize the “need-to-know” implicit in the request. Louis Harris and Associates asked a sample of Americans in 1983 whether they felt that certain government organizations that collect or use information “limit their personal information about individuals to what is really necessary or whether they ask for too much information.” Twenty-four percent of the total public believed the Census Bureau did not need to collect all the personal information it requested. Thirty-nine percent of the congressional representatives and their staff contacted felt the same way (Harris (1984)). These responses may reflect a broader, related belief that governments – including U.S. Federal statistical agencies – have no “right-to-know” this information. Those who have an Orwellian fear of government as “big brother” may not be satisfied with a demonstration of the government’s need to know this information. Indeed, such an effort to demonstrate the usefulness of this information to the government may exacerbate the perceived threat to confidentiality. The U.S. Census Bureau can attempt to distinguish itself as somehow separate and apart from the rest of the U.S. Federal government, but this effort is difficult. From the outside, the Federal government may appear to be monolithic. Many individuals assume that what one part of the government learns is known by all parts of the government. A fairly significant proportion of the population assumes that the Census Bureau shares data on individuals with other government agencies (Baxter (1984)). The Census
Bureau's general policies to the contrary are not widely known or believed. The respondent who fears that the government knows too much about him/her is not likely to draw a distinction between the Census Bureau and the rest of the government (see e.g., Harris (1984)).

4.2. Fear of Computers

Another perceived threat to confidentiality which information collectors face is rooted in "computer phobia." This paper has already discussed in some detail some of the ways in which automated data processing techniques can improve the U.S. Census Bureau's operations. Direct threats to confidentiality are posed by computers, but the Census Bureau is finding solutions to these direct threats. It is more difficult to deal with perceived threats to confidentiality posed by the increasing use of computers.

Surveys have demonstrated that a significant number of people are uninformed and suspicious about computers (Harris (1984)). Part of the phobia about computers is based on a belief that computers can be superefficient (Kagay et al. (1984)). Some individuals believe that enormous amounts of seemingly trivial data are stored and easily combined to produce shockingly detailed descriptions of an individual's characteristics, habits and lifestyle. Computer phobia also may be rooted in a belief that existing computer security techniques are inadequate. Some individuals believe that computers can be "tapped," or entered by unauthorized persons, almost at will, and that no data kept in a computer file are safe from unauthorized disclosure (see e.g., Harris (1984)). How can the U.S. Census Bureau deal with the problems posed by these perceived threats? Given that many of these perceptions arise from incomplete or misinformation, the Census Bureau may be able to address many of these concerns through its public information and outreach activities. For instance, the Census Bureau could publicize the fact that it does not collect social security numbers as part of the decennial census. Furthermore, the Census Bureau always strives to demonstrate and communicate the need for the information it requests. More could be learned, however, about the most effective ways to educate the respondent about the purposes and uses of the data collected.

4.3. Effects of Confidentiality Assurances

Additional research on the effects of confidentiality assurances also is indicated. What are the consequences of providing full and accurate information about the U.S. Census Bureau's confidentiality policies and practices? Does the dissemination of this information give rise to perceived threats regarding the confidentiality of the Census Bureau's data? Perhaps more importantly, does dissemination of information have an effect on the public's willingness to cooperate with census and survey activities? Census Bureau personnel have traditionally assumed that successes in both the mandatory decennial census and its voluntary surveys are ultimately dependent upon a sense of public trust which encourages support for and cooperation with its efforts. Yet a systematic review of existing social science research suggests that the effect of privacy and confidentiality assurances on willingness to participate in survey research activities may not be as strong nor as direct as it is generally assumed.

Beginning in the 1960s and continuing to the present day, social science researchers have been concerned with the implications of privacy and confidentiality issues for social

---

3 A number used for administrative purposes, primarily to distribute Social Security pensions. These numbers are not equivalent to personal identity numbers since there are individuals who do not have one.
(particularly survey) research, (see e.g., Carlson (1967), American Statistical Association in Bulmer, ed. (1979), National Research Council (1979)). One manifestation of this concern was a spate of research examining the actual effects of privacy and confidentiality upon social research. The largely counterintuitive result of these studies was the absence of empirical evidence of a strong, unambiguous relationship between (generally high levels of professed) concern with privacy and confidentiality issues, and response rate and quality.

(Given the assumed strong, positive relationship between perceptions of the confidentiality of individually identifiable data and response behavior, this finding is counterintuitive. Given the findings of the wealth of social-psychological literature which addresses the attitude-behavior relationship, however, the absence of such a linkage is not surprising.)

4.4. Some Experimental Studies

Examinations of the impact of privacy and confidentiality upon social research generally fall into one of three categories involving experimental examinations of the impact of (1) confidentiality pledges, (2) interview privacy conditions on response rates and quality, and (3) interviewers’ perceptions of the extent to which respondents’ concerns with privacy and confidentiality contribute to response difficulties. In a 1973 review article Streib discusses the effect of interview privacy (“a situation in which the interviewer and respondent are able to converse without being heard by third persons who might influence the responses,” Streib (1973, p. 276)) on response behavior in research on families. While he uncovers some conflicting opinions about the effect of privacy on the quality of response (see e.g., Pearlin (1971), Blood and Wolfe (1960), Taietz (1962)), Streib does identify some consensus that responses to factual questions are not influenced by interview privacy. Attitude and perceptual questions, on the other hand, do appear to be influenced by interview privacy. Streib reports no evidence, however, that interview privacy affects willingness to be interviewed.

Unlike investigations of interview privacy (which actually vary interview conditions), examinations of the effects of confidentiality pledges look for variations in response based on promised future uses of the information that the respondent provides. Thus the issue of trust is implicit in this research. Studying the effects of confidentiality pledges on response, item nonresponse, and data quality, Singer (1978) found that confidentiality pledges did not have a significant effect on item nonresponse.

“Even though it was impossible to discern any effect of confidentiality on overall response rate to the interview, assuring respondents of absolute confidentiality had a small but consistent effect on willingness to answer individual questions. Nonresponse rates for sensitive questions were consistently and sometimes significantly lower when people were told that their replies would be held in confidence. Furthermore, there is at least the suggestion that a promise of confidentiality enhances the quality of response to the most sensitive items on the interview. Thus, if respondents can be assured of the confidentiality of their replies, responses to sensitive questions will benefit” (Singer (1978, p. 159)).

A complementary study is Singer and Kohnke-Aquire’s examination of interviewer expectation effects. They found, as they had expected, that “promising confidentiality was seen by interviewers as enhancing both the quality of response and people’s willingness to participate,” (Singer and Kohnke-Aquire (1979, p. 249)). Response behaviors did not meet these interviewer expectations, however. Singer and Kohnke-Aquire found no discernable effect of these interviewer expectations on overall response rate. Other evidence may support this link, however.

In 1979 the Panel on Privacy and Confidentiality as Factors in Survey Response reported
the results of a survey designed to "test a technique for assessing response behavior to census-type inquiries under alternative degrees of assurance of confidentiality," (National Research Council (1979, p. 91)). Once again, disparity among the confidentiality pledges produced little patterned variation in response behaviors.

"Overall response to the census-type questions across all treatments was high; more than 91 percent. The experiment did not identify promise of confidentiality as the major cause for nonresponse. It did reveal that there are apparently some people for whom concern about lack of confidentiality contributes to nonresponse. A monotonic increase in refusal rates associated with treatments was found as the degree of assurance of confidentiality decreased, ranging from 1.8 percent for the strongest assurance to 2.8 percent for denial of confidentiality. A similar pattern was observed for total noninterview rates from all causes, ranging from 7.2 percent for the group with the strongest to 10.0 percent for the group with no assurance" (National Research Council (1979, p. 116)).

The Panel's efforts to uncover additional empirical research in this area by informally surveying data-collecting organizations were not encouraging. The Panel found "little hard evidence, based on empirical tests under controlled conditions, that an assurance of confidentiality improved response levels," (National Research Council (1979, p. 157)). The 33 individuals (representing 28 organizations) who responded to the Panel's request for information provided little evidence that either the Privacy Act of 1974 or the Freedom of Information Act had affected response rates, although some of the respondents "expressed concern that these acts might have a future impact," (National Research Council (1979, pp. 141–142)). Thus, while the survey revealed a general consensus among data-gatherers that would-be respondents fear data disclosure, and that confidentiality pledges elicit better responses, it shed no new light on the actual relationship between concerns with confidentiality and response behaviors.

Meyers and Oliver (1978) and Olson and Klein (1980) report on interviewers' perceptions of the extent to which respondents' privacy and confidentiality concerns contributed to response problems in two nonexperimental situations. Meyers and Oliver focus on the impact of public concern with privacy (and anti-government sentiment) on refusal to participate in the National Medical Care Expenditure Survey. A noninterview report form was used to obtain interviewer comments on reasons for refusals. Meyers and Oliver rank-ordered the resulting eleven dimensions of refusal for 985 noninterviews, finding that objection to government invasion of privacy was the least frequently cited reason, and desire to protect one's privacy was the sixth most important reason (lack of interest, unspecified reasons, and lack of time being the three most common reasons). Moreover, they found that concern with privacy remained a relatively unimportant reason for refusal to participate when the data were examined by race, sex, place of residence, and age. Olson and Klein explore dimensions of refusal for the Income Survey Development Program (a survey sponsored by the U.S. Department of Health and Human Services). Like Meyers and Oliver they also based their study on a refusal form completed by interviewers. They found invasion of privacy to be one of the most common reasons for refusal (Olson and Klein (1980, p. 553)). More importantly, they report "if we leave the 'other' category aside, we observe that 'don't want to spend time' and 'invasion of privacy,' the most common reasons for refusal, are in many instances four or five times as frequent as alternative categories," (Olson and Klein (1980, p. 553)). Like Meyers and Oliver, they find that when the data are tabulated by region, age, household size, gender, and race, "no one demographic group appears to have a greater tendency than another group to refuse
because of privacy concern,” (Olson and Klein (1980, p. 554)).

More recently, Singer “reports on a survey of public attitudes toward issues of informed consent in social research” (Singer (1984, p. 501)). She found “a high level of expressed concern about such matters as confidentiality of response and study sponsorship” and “significant interrelationships between the expression of such concerns and willingness to answer sensitive questions on the interview, though generally such relationships are small” (Singer (1984, p. 501)).

Summarizing the results of these studies, the evidence suggests that the effects of interview privacy on response rate and quality are overemphasized. There is little evidence that interview privacy affects willingness to be interviewed or responses to factual questions, though there is some suggestion that it affects responses to attitude and perceptual questions (Streib (1973)). We have evidence that both interviewers and members of the public expect promises of confidentiality to improve response rate and/or quality (Singer and Kohnke-Aquire (1979), Singer (1984)). In keeping with these expectations, one study found that for some people concerns with confidentiality seem to contribute to nonresponse, and another found a significant effect of confidentiality pledges on item nonresponse with some suggestion of an effect on response quality for particularly sensitive items (National Research Council (1979), Singer (1978)). On the other hand, these same two studies agree that confidentiality has no major effect on overall response.

To complicate our understanding of the effects of confidentiality on respondents even further, other research suggests that in some cases confidentiality assurances may provoke negative reactions. Reamer suggests that merely broaching the subjects of confidentiality and anonymity (which otherwise might not have been of concern to the subjects) may actually increase apprehension, while having no effect on response rate or quality (Reamer (1979, pp. 503–504)). Additional evidence demonstrates that on occasion, privacy and confidentiality concerns have, in fact, seriously impeded or threatened to endanger survey efforts, (see, e.g., Catterall (1983), Pond (1983), Josephson (1970)). The combination of counterintuitive and contradictory findings prompts us to search for some underlying pattern that will help us identify the conditions under which privacy and confidentiality concerns will become important factors in the conduct of social research.

4.5. The Attitude-behavior Relationship

Research suggests that one explanation for the absence of a consistent reflection of concerns with privacy and confidentiality in response behavior is the complex nature of attitude holding. Boruch points out that “if we look over the ten or twenty controversial cases in which it has been announced that privacy has been a major factor in the disruption of a survey or a large-scale program evaluation, we see that it is typical that the privacy issue is confounded almost inextricably with a number of other issues,” (cited in Merriam et al. (1980, p. 435)). Westin and Baker’s review of poll results suggests that:

“underneath the concern over confidentiality which many people have, and the solid minority concern about ‘invasion of privacy’ as a general theme, lie opinions which are at least as varied and at some points reflect inner conflict on the part of the respondents. ... Some of these differences on specific items reflect the fact that there are real dilemmas present in setting the balance between the information needs of organizations and the rights of individuals” (Westin and Baker (1972, p. 476)).

In short, “public opinion is almost invariably composed of a very complicated mixture of
conflicting and contradictory attitudes and is far from being the simple entity so often declared categorically to favour this course or to be outraged by that," (Social and Community Planning Research Working Party in Bulmer, ed. (1979, p. 71)). It is important to remember that attitudes may not be well-formed, well-informed, or consistent over time, much less across individuals (or groups of individuals). Under these conditions, it is not surprising to find a sporadic attitude-behavior relationship. A second factor which may contribute to the absence of the expected relationship is the difficulty encountered in obtaining accurate information on attitudes and opinions. We know that respondents may express opinions on subjects on which they have no information or previous knowledge. We also know that poorly designed survey instruments can lead or frighten respondents, thus biasing responses. We assume that individuals occasionally respond with "socially desirable" rather than actual opinion information. Regardless of whether respondents' misreporting of attitude information is intentional or unintentional, the result is the same: distortions that lead to expectations of behavior based on mismeasurement of public opinion.

A third explanation for the absence of the expected relationship highlights the differences between attitude holding and behavior based on attitudes. Fromkin suggests that where confidentiality is concerned, perceived benefits from providing personal information may also tip the balance of conflicting attitudes. Fromkin found that college students were "more responsive to personal questions" if they felt that providing the information would be helpful to them personally, e.g., if it would affect their financial aid, (Fromkin (1976)).

This suggests a reason why a public that appears to be concerned with the confidential nature of information held by employers, creditors, insurers, etc., nevertheless routinely provides them with personal information. Without engaging in a debate on the nature of the attitude-behavior relationship, or delving into a discussion of individuals' decision-making processes, we must consider the possibilities that: (a) individuals frequently act in response to "scripts" (e.g., Abelson (1981)) rather than on the basis of attitudes and rational decision processes; and (b) even when actions are influenced by concerns, attitudes, or opinions, the presence of conflicting concerns may result in unexpected or poorly predicted behavior. One implication is that individuals perform an internal benefit-cost analysis in deciding whether or not to provide personal information, based on such factors as previous experience with invasions of privacy, probable benefits, and anticipated costs. Social science literature argues that, in fact, most individuals do not consistently nor routinely employ such formalized thought processes; but the underlying notion of balancing benefits and costs may be relevant in those instances in which the individual's attention becomes focused on one aspect (benefit or cost) of an issue. What is needed is an attempt to define under what conditions, and for what types of respondents, privacy and confidentiality issues are both more salient, and more likely to result in non-cooperative behavior than others. For example, under what conditions are confidentiality concerns more or less salient than the issue of respondent burden? When do anticipated benefits from providing a response outweigh feelings that a particular question is an invasion of privacy, or fears that the confidential information provided might be shared?

4.6. Suggestions for Additional Research

These tentative explanations suggest three areas for additional research on public perceptions, each of which is relevant to the issue of
confidentiality but which, in addition, has broader applications. While the primary benefits of such research to the U.S. Census Bureau would clearly be applications to planning, decision-making, and outreach activities, it offers an exciting opportunity to contribute to academic research currently underway in these or related areas. The first item on this research agenda is the development of better means of assessing opinions, (e.g., combining content analyses of documents and surveys of opinion leaders with national opinion surveys in hopes of triangulating upon a better assessment of actual opinion structures). A second area which is ripe for additional research is the nature of the attitude-behavior relationship. The third promising research option is the development of mechanisms for anticipating changes in climates of opinion. In considering the implementation of new census-taking, data processing and information dissemination procedures, our goal is to anticipate possible negative reactions and thus avoid them through modified decision plans. Such proactive decision-making would be enhanced by an ability to understand (and anticipate) how issues emerge and become defined as “problems” within various segments of the public. Current investigations which take such an “agenda-setting” approach to opinion formation suggest that focusing events, issue linkages, intersecting or reinforcing issue climates, and groups may play important roles in defining and increasing the salience of issues (see e.g., Kingdon (1984), McDonald (1984), Cook in Lewis, ed. (1981)).

In addition to examining how such social science theory may be extended and applied, it is desirable to consider ways in which issues management, futures forecasting, and other “early warning system” techniques and processes may be applied by public agencies. Research on issue emergence and the opinion formation process would not only prove invaluable to the U.S. Census Bureau (and other organizational) decision-making, but also would provide an opportunity to make significant contributions to scholarly research currently being undertaken to address some of these concerns (particularly in the interdisciplinary field of policy analysis).

5. Some Concluding Thoughts

5.1. Indirect Threats to Confidentiality

Before summarizing our research recommendations it is worth mentioning one other, more philosophical, issue raised by the U.S. Census Bureau’s efforts to maintain the confidentiality of each respondent’s answers. The Census Bureau is faced with a difficult problem as it tries to deal with what can be called “indirect” threats in maintaining control and access to data. These indirect threats are caused not by the Census Bureau’s actions, but rather by others who use or manipulate the Census Bureau’s data in ways that invade the privacy of individuals or segments of society. For example, as detailed in Section 2, someone may attempt to decipher a disclosure avoidance scheme used by the Census Bureau to identify confidential information supplied by a particular respondent. To do this he/she might perform arithmetic operations on related data tables to draw inferences that could allow identification of a person with certain unique characteristics (for instance, someone who lives in a certain block, with a certain income). However, because of the lack of certainty based on these inferences, this threat must be regarded as indirect.

How far should the U.S. Census Bureau go in attempting to avoid disclosure due to the persistence of data users? The Census Bureau currently provides for the suppression of some data because, in combination with other Census Bureau supplied data, such inferences may be possible. In other instances the Census Bureau does not perform complementary suppression because the risks of disclosure are
considered too low. Nevertheless, the mind of a determined and talented user could break many complementary data suppression techniques given enough time and computer resources. Therefore, the Census Bureau must assess the resiliency of its disclosure avoidance methods and the likelihood and potential harm of invasion attempts.

Another indirect threat to confidentiality is presented by those who combine census data with data from other sources to form detailed demographic and economic profiles of a small geographic area. One such mixer of census data with other data is the well-publicized “Geodemographic” data base (Burnham (1983)). A group of political and marketing consultants have used publicly available census data to categorize postal ZIP codes by income and other characteristics to target particular political or advertising messages. ZIP codes are described by phrases such as “Blue Blood Estates,” “Bohemian Mix,” “Furs and Station Wagon,” “Bunker’s Neighbors,” “Urban Renewal,” and “Tobacco Roads.” The consultants developed forty such groupings and assigned each of 36 000 ZIP codes to one of the forty categories based on census data.

Whether such a political or marketing device as Geodemographics is an invasion of privacy or not, the U.S. Census Bureau faces the problem that its data, normally used as an instrument of good, could be used for less noble purposes. The admission that knowledge and information can serve both good and bad purposes should not be startling. The practical implications for the Census Bureau, however, are not self-evident. The most obvious way in which the U.S. Census Bureau addresses this type of indirect threat is in the process followed to decide what questions to ask on its censuses and surveys. The content of its questionnaires is reviewed in many places (the Office of the Secretary of Commerce, the Office of Management and Budget, and in some cases, Congress). Sensitive questions like religious affiliation are not asked. Some other sensitive questions like income have been asked in the past but only on the census long form sent to a sample of the population or smaller current surveys. By excluding the income question from the census short form sent to everyone, one cannot make valid statistical measures of income for very small geographic areas like individual blocks. They can be made for larger geographic areas like census tracts that produce less precise descriptions of where the wealthy live. (It should be noted, however, that other data collected from all census respondents, such as the value of the house, can be used as a substitute for income at the block level.) For the decennial census the decisions about what questions to ask are made through an established process, one that has great legitimacy and involves significant consultation with other agencies, Congress and non-governmental organizations and individuals. Once that process is conducted and the question is placed on a questionnaire, the Census Bureau’s general policy is that tabulation and publication of that data should follow (with appropriate disclosure avoidance techniques applied and caveats as to data quality noted).

The U.S. Census Bureau has to be aware of indirect threats to the American public’s right of privacy. It has to structure its data products so that the intentional defeat of its disclosure avoidance techniques is very difficult. The Census Bureau must strike the proper balance between the risk of disclosure and the need to publish useful data products. It also has to keep in mind the potential misuse of its data when it makes recommendations to the Executive Branch and Congress for new census questions or is paid by another agency to collect certain data in a survey.

5.2. Summary of Suggestions for Future Research
The preceding sections have illustrated the effects which privacy and confidentiality issues
have on the work of the U.S. Census Bureau. They have pointed out various opportunities to improve the Census Bureau’s understanding of the confidentiality issue and to maintain the confidential nature of the data collected in the face of new technologies. Three areas which are particularly ripe for additional research are: (1) techniques to protect and maintain the confidentiality of the data which the Census Bureau collects, processes, stores and disseminates; (2) public concerns with privacy and confidentiality issues generally, and related to the Census Bureau specifically (e.g., perceptions of the census as a potential invasion of privacy); and (3) the linkage between these public perceptions and behavior (i.e., the potential effect of privacy and confidentiality concerns upon the public’s willingness to cooperate with censuses and surveys). For the Census Bureau, the broad objective of research on disclosure avoidance in frequency counts is to identify and develop a single, preferred model of disclosure avoidance – or to demonstrate that narrowing the choices to one method is unwise. In the area of cell suppression methodology, alternative linear models of specialized mathematical programming models and algorithms which are efficient for large problems are needed; the work of Gusfield (1984) must be evaluated and extended. Further research needed in micro-data disclosure includes: investigating the applicability of controlled rounding; exploring the effects of sample design or selection on preserving respondent confidentiality; top-coding and the extent to which the distribution of data in topcoded categories can be made available; further study of microaggregation techniques; and the problems of setting suitable population thresholds. A major, virtually unexplored area of research for all three data types is to quantify the effects of various disclosure avoidance methods upon data use. The seminal work of Paass (1985) must be evaluated for Census Bureau microdata and extended to provide adequate disclosure protection.

New uses of emerging computer and telecommunication technologies call for indepth analyses of the security vulnerabilities of automation configurations and the development of guidelines for implementation of the necessary security measures. In addition to actually protecting the confidentiality of the data it collects, however, the U.S. Census Bureau needs to preserve the public perception that it is successfully doing so. This argues for a research agenda to address the issue of public perceptions, including: the development of better means of assessing opinions; additional research on the information-attitude-behavior relationship; and the development of mechanisms for anticipating changes in climates of opinion to allow for proactive decision-making.

The U.S. Census Bureau hopes that by combining and sharing its ideas with those of other data collectors and providers, as well as data users and members of the academic community, it can support the integrated development of theoretical and applied research on these complex topics. The Census Bureau encourages others to join its continuing efforts to improve the understanding of all confidentiality related issues.

6. References


McDonald, S.-K. (1984): Power Over Decision: Agenda-Setting in the U.S. House of Representatives. Published doctoral disser-
tation, Northwestern University, Evanston, Illinois.

Received December 1985
Revised April 1986