Journal of Official Statistics, Vol. 23, No. 3, 2007, pp. 299–317

Optimizing the Use of Microdata: An Overview of the Issues

Julia Lane[1]

"It is becoming clear that advances in technology and increased use of administrative records may, at some point in the future, render our current disclosure avoidance procedures inadequate. At the same time the . . . federal statistical system face[s] increasing demands for more, better and more recent data to meet critically important public policy and research needs."[2]

"The extraordinary growth of electronic infrastructure, capacity, and use in the past decade has posed a profound new set of questions about the control, dissemination, power and use of information. On the one hand the high speed internet and the World Wide Web, email, electronic shopping, and cell-phone use have opened up extraordinary new worlds of communication and are changing the way we work, play, and learn. On the other, as the electronic world enters our daily lives, the private space untouched by the intrusions of cyberspace and information seekers shrinks – for individuals, firms, and organizations. . . . There is also another challenge. The need to build more efficient surveillance networks to combat potential terrorist attack argues for less privacy for the individual person or firm to guarantee the security of the society in general. It is in this environment that citizens, business and technology leaders, and policy makers have to figure out how to understand, manage, and regulate the new cyber world."[3]

Key words: Confidentiality; microdata access; access modalities; risk/utility tradeoff.

1. Introduction

New capacities to collect and integrate data offer expanded potential for scientists and policy-makers to understand factors contributing to key national priorities – like job, income and wealth creation, as well as career path and retirement decisions made by individuals. This capacity can also contribute to meeting a critical national security need. The major security threat to the United States is inherently human and an improved ability

[1] NORC at the University of Chicago, 1155 East 60th Street, Chicago, IL 60637-2799, U.S.A. Email: lane-julia@norc.uchicago.edu

[2] Pat Doyle, Maintaining Data Access and Dissemination in Light of Changes in Technology. Draft document, August 9, 2002.

[3] Margo Anderson, SBE/CISE workshop, March 15-16, 2005.

Acknowledgment: This article was written as part of a session organized to honor Pat Doyle at the 2005 American Statistical Association meetings in Minneapolis. Many ideas are derived from discussions with John Abowd, Pat Doyle and Laura Zayatz. Thanks also to Nick Greenia for extensive discussions on data quality and harm issues, Bill Winkler for alerting me to additional data quality and confidentiality literature, Miriam Heller and Sang Kim for their ideas about the relationship between cyber infrastructure and confidentiality, Nancy Lutz for her help in developing the model and Guy Almes, Fredrik Andersson, Matt Freedman, Cheryl Eavey, Nancy Gordon, and Dan Weinberg for their suggestions.

to understand and predict malevolent behaviors can provide one means for addressing that threat.

Two substantial challenges face collectors and producers of social science data as a result of this increased capacity. The first is how can the information derived from vast streams of data on human beings be used while protecting confidentiality? The second is the essence of good science: how can society best provide and promote access to rich and sensitive data so that empirical results can be generalized and replicated?

An existing community has already focused on protecting confidentiality. In particular, U.S. federal statistical agencies have devoted substantial resources to both statistical and technical ways to protect confidentiality.[4] The Social and Behavioral Research Working Group recently drafted a report entitled "Achieving Effective Human Subjects Protection and Rigorous Social and Behavioral Research" for the Human Subjects Research Subcommittee of the Committee on Science, National Science and Technology Council, PITAC[5] recently issued a report on cyber security that addressed some confidentiality issues, and numerous studies have been undertaken by the National Academy of Sciences and the Committee on National Statistics. The National Science Foundation has also been active in the area of cyber trust, and the PORTIA (Privacy, Obligation and Rights in Technologies of Information Assessment) project based at both Yale and Stanford universities directly addresses many of the key issues.

However, focusing on confidentiality protection alone is likely to lead to piecemeal approaches and result in outcomes that are in the best interests neither of decision-makers nor of society at large. The appropriate approach is to optimize the amount of data access, subject to meeting key confidentiality constraints. And, although Fienberg et al. (2004) and Duncan et al. (2001, 2003, 2004), in particular, have been vocal advocates of preserving statistical utility of tabular data, only recently is attention being paid by the statistical community to optimizing access to microdata.[6]

This article begins by discussing current confidentiality protection techniques for public use microdata files accompanied by illustrations of some consequences for the typical type of analyses performed by economists. It then describes the challenges that are emerging as a result of technological advances, reviews alternative access modalities and develops a simple economic framework. The article concludes with a suggested research agenda.

2. An Overview of Current Confidentiality Protection Techniques and Their Consequences

A good description of the practical application of microdata disclosure limitation techniques practiced at the U.S. Census Bureau is provided in Zayatz (2005). She points

[4] Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, Pat Doyle, Julia Lane, Laura Zayatz and Jules Theeuwes, North Holland, 2001.

[5] President's Information Technology Advisory Committee.

[6] Final guidelines have, however, recently been released by the United Nations Economic Commission for Europe (25 October 2006) "Managing Statistical Confidentiality and Microdata Access" http://www.unece.org/stats/documents/tfcm/1.e.pdf and an entire session at a recent UNECE/Eurostat conference was devoted to the topic (Monographs of official statistics, Work session on statistical data confidentiliaty, Geneva 9-11 November 2005, ISBM 92-79-01108-01). See also Abowd and Lane (2003).

out that the risk of disclosure can be reduced either by reducing the amount of information or by perturbing the data.

The means used in reducing the amount of information include variable deletion, recoding categorical variables into larger categories (perhaps using thresholds), recoding continuous variables into categories, rounding continuous variables, using top and bottom codes, using local suppression and enlarging geographic areas. Data can be perturbed by means of noise addition, record swapping, rank swapping, blanking and imputation, micro-aggregation or by multiple imputation/modeling to generate synthetic data.

Although each of these approaches can have an effect on the validity of social science analysis, the decision to apply them does not fully capture the costs to society of reduced data quality. A good discussion of the issues is provided in Smith (1991). The effect of decisions on top coding is well summarized in the earnings inequality literature. Burkhauser et al. (2004) found that changes in top coding rules of one of the most important public use surveys, the Current Population Survey (CPS), in the 1990's artificially *increased* measured earnings inequality.

The key problem is that a standard measure for calculating earnings inequality is the Gini coefficient which ranges between 0 and 1. A value of 0 corresponds to a situation where everyone has the same income, or perfect equality. The value of the coefficient increases as the richest percentiles in society earn higher proportions of income. Top coding artificially reduces the maximum income level, resulting in a coefficient that is biased down. Arbitrary changes in top codes can change the Gini coefficient up or down – artificially changing earnings inequality.

As Mishel and Bernstein point out in a debate between Robert Lerman (1997) and Jared Bernstein and Lawrence Mishel (1997):

> However, before we can reliably measure inequality trends in the CPS or, for that matter, any other public-use data set, we must deal with the issue of top codes, an issue that becomes particularly germane when earnings at the top are growing quickly relative to those elsewhere in the earnings distribution. . . .There are a number of ways to approach the top-coding problem. One is simply to ignore top coding. Doing this, however, is a problem in Gini analysis, because nominal wage growth over a period when the top code does not change or increases only slightly will lead to increasing shares of earners who are top coded, thus biasing the Gini coefficients downward. Such a downward bias applied between 1981 and 1987, when the top code stayed between $75,000 and $99,999, before doubling in 1988 (pp. 3–4).

A clear illustration of the consequences is provided by the graph reproduced from Burkhauser et al. (2004) below. The bottom line in the graph shows that had top coding on the Current Population Survey been consistent, then earnings inequality, as captured by the Gini coefficient, would have increased steadily between 1975 and 2001. However, top coding did not remain consistent. The second line plots the Gini coefficient derived from public use files. The public-use top code was $99,999 until 1995 when the U.S. Census Bureau both raised the public-use top code to $150,000 and assigned cell means for persons with earnings above the top code. The surge in earnings inequality from about .34 to .39 is completely an artifact of that top coding decision. It is worth noting that the 1993 surge in the third line reflects a data collection, rather than a reporting decision. In that year, the U.S. Census Bureau changed its internal system to permit the recording of incomes of \$999,999, rather than \$249,000 (between 1979 and 1984, the maximum permissible was \$99,999).

In sum, the approach used to top coding public-use data, as well as internal administrative decisions, can result in vastly different information being provided to policy makers. And, to repeat the theme of the article, inappropriate action by the policy-makers can result in outcomes that are neither in the best interests of decision-makers nor of society at large.

The effect of top coding on other standard uses of public-use files is also very clear, since the theory concerning regressions when the dependent variables are censored from both above and below is quite well developed. Indeed, the 2000 Nobel Prize was given, in part, to Jim Heckman for his path-breaking work on statistical approaches to dealing with the econometric problems posed by selective samples.[7]

A brief example using an earnings regression model illustrates the effect on regression coefficients. Suppose we have an earnings regression model

$$Y_i = X_i'\beta + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma^2)$$

where Y_i is the earnings of individual i, and X_i is a set of that individual's characteristics, but the model is censored from below by a and above by b. It is straightforward to show that standard least squares regression will result in slope coefficients that are downwardly biased. Consistent coefficients can be estimated if the distribution of the error term given the regressors is known, and some of Heckman's most important work has dealt with doing just this by modeling the behavioral decision that leads to censoring from below. The fundamental problem with arbitrary top coding is that the distribution is not provided, making it extremely difficult to recover consistent estimates. Although several alternative estimators have been developed, using different assumptions about the distribution underlying the top coded values, there is still wide divergence among estimated coefficients.

The implications of this divergence can be quite substantial. Two key issues of interest to policy-makers are the black/white earnings gap and the return to education. The following table, which is reproduced from Chay and Powell (2001), illustrates the wide divergence in estimates using different techniques (using top coded Social Security data). The first two columns (OLS1 and OLS2) use ordinary least squares approaches that do not attempt to address the distributional consequences of top coding. Maximum Likelihood Estimation (MLE) procedures, the results of which are reported in Column 3, assume that the errors are normally distributed and homoskedastic. Chay and Powell develop three semi-parametric estimators for censored regressions: the CLAD (Censored Least Absolute Deviations), the SCLS (Symmetrically Censored Least Squares), and the ICLAD (Identically Censored Least Absolute Deviations).

The first panel reflects the results of running standard earnings regressions that estimate the effect of race on the log of earnings using these six different approaches. The coefficients

[7] To quote from the Nobel Prize Committee "Available microdata often entail **selective samples**. Data on wages, for instance, cannot be sampled randomly if only individuals with certain characteristics – unobservable to the researcher – choose to work or engage in education. If such selection is not taken into account, statistical estimation of economic relationships yields biased results. Heckman has developed statistical methods of handling selective samples in an appropriate way." http://nobelprize.org/nobel_prizes/economics/laureates/2000/press.html

reported in each column can be approximately interpreted as the percentage difference in earnings between blacks and whites in each year, controlling for age. Briefly, not only do estimates of the black/white earnings gap range from .35 to .63 log points in 1963, but estimates of the degree to which the gap closed between 1963 and 1971 range from .06 log points in the black/white earnings gap using OLS regression techniques to .15 log points using alternative measures. Policy-makers might look at one set of numbers and conclude that the racial earnings gap was closing rapidly; at another set and conclude that it was closing slowly. In the former case, the policy-maker might well conclude that no intervention was required; in the latter, that intervention was necessary. One of those decisions would be wrong, although it is not clear which is the incorrect decision. Certainly one would be neither in the best interests of decision-makers nor of society at large.

The second panel reflects the results of using the different estimation techniques to calculate the return to education – another topic of key interest to policy-makers. A policy-maker who only used information from the second column would note that the returns to education had gone from about 1% in 1963 to approximately zero in 1973, and would be forgiven for concluding that further investment in education was unnecessary. A policy-maker examining the final column would see that the return to education was a consistent 7%, and could conclude that further investment would be a wise allocation of public monies (Table 1).

The consequences of the other disclosure limitation techniques – such as recoding, rounding and data swapping – are less well documented, although each should act to bias coefficients towards zero. It is remarkable, however, that despite the fact that statistical agencies publish extensive and high-quality documentation that informs users of the consequences of different sampling procedures and nonsampling errors, and how to adjust estimates accordingly,[8] the effort to achieve disclosure limitation is hampered by concerns that such information would permit researchers to "back out" the disclosure limitation algorithms. It would seem clearly preferable that the holder and producer of microdata should list specific limitations that affect the ability of the microdata to support valid analyses. Alternatively, the data producer should either provide access to suitable microdata so that users can determine which types of estimation procedures to use, or provide suitable auxiliary information with public use microdata so as to permit the approximate reproduction of the results that might be obtained on the original microdata. The remote access approaches being used by a number of European agencies represent promising moves in this direction.

Part of the challenge is that social scientists use microdata in many different ways and it is difficult to directly define what is meant by data quality. An illustrative example is provided by the workshop on total survey error that the National Institute of Statistical Sciences (NISS) held in March 2005, from which it is clear that quality concepts are difficult to use in most specific settings.[9] The Eurostat definitions, which lack metrics, are (1) relevance, (2) accuracy, (3) timeliness, (4) accessibility and clarity of results, (5) comparability, (6) coherence, and (7) completeness (Haworth et al., 2001). Winkler (2005e) has provided some metrics to diagnose serious problems with a file, but these do not assure analytic quality. As Winkler (2005f) has pointed out, the challenge when it

[8] A good example is the 228 page document (U.S. Department of Labor 2002) on the design and methodology of the Current Population Survey.

[9] I am grateful to Bill Winkler for providing me with the workshop information.

Table 1. Estimated effect of race and education on log-earnings (estimated standard errors in parentheses)

	OLS1	OLS2	MLE	CLAD	SCLS	ICLAD
Black-White Gap						
1963	−0.355 (0.033)	−0.183 (0.038)	−0.629 (0.044)	−0.416 (0.027)	−0.444 (0.031)	−0.474 (0.032)
1964	−0.349 (0.032)	−0.154 (0.038)	−0.674 (0.044)	−0.428 (0.033)	−0.444 (0.036)	−0.473 (0.031)
1970	−0.262 (0.032)	−0.115 (0.037)	−0.508 (0.044)	−0.278 (0.020)	−0.302 (0.031)	−0.338 (0.029)
1971	−0.242 (0.031)	−0.111 (0.038)	−0.486 (0.044)	−0.244 (0.022)	−0.287 (0.032)	−0.312 (0.031)
Returns of education						
1963	0.041 (0.003)	0.012 (0.004)	0.102 (0.004)	0.051 (0.004)	0.068 (0.007)	0.073 (0.003)
1964	0.040 (0.003)	0.013 (0.005)	0.103 (0.004)	0.064 (0.006)	0.079 (0.007)	0.075 (0.003)
1970	0.037 (0.003)	0.003 (0.005)	0.101 (0.004)	0.055 (0.003)	0.066 (0.006)	0.071 (0.003)
1971	0.035 (0.002)	0.002 (0.004)	0.100 (0.004)	0.054 (0.003)	0.065 (0.005)	0.070 (0.003)

Notes: The dependent variable is the natural logarithm of annual taxable earnings. Regressions also include a constant, and age and age-squared as explanatory variables. Observations with nonpositive earnings are dropped from the analysis. The sample sizes for 1963, 1964, 1970, and 1971 are 8525, 8529, 8391, and 8275, respectively. The OLS2 specification also drops top-coded observations, leading to sample sizes of 4632, 4267, 4485, and 4163. MLE is Tobitt maximum likelihood; CLAD is censored least absolute deviations. SCLS is symmetrically censored least squares; ICLAD is identically censored least absolute deviations.

comes to maintaining quality in a masked file is constituted by the fact that certain aggregates such as higher order moments must be accurate (say for regressions).

3. Future Data Collections and the Associated Confidentiality Challenges

The previous section demonstrated that current statistical disclosure techniques act in unknown ways to severely diminish the utility of microdata for analysis. It is also clear that the challenge to protecting the confidentiality of microdata will only increase. In addition to the challenges posed by the increased capacity for disclosure, that were thoroughly documented in Doyle et al. (2001), new data collection modalities are emerging that pose much greater likelihood of disclosure, and there is much greater access to administrative data.

Although data collection on individuals and organizations has historically consisted of either survey-based or administrative data, cyber infrastructure[10] advances have fundamentally changed the way in which scientists are collecting information and modeling human behavior. Indeed, a recent National Science Foundation solicitation, entitled "Next Generation Cyber Tools" noted that new ways have been developed to improve both domain-specific and general-purpose tools to analyze and visualize scientific data – such as improving processing power, enhanced interoperability of data from different sources, data mining, data integration, and information indexing.[11] A calculation at the recent NSF-supported workshop[12] of how many terabytes of data would be necessary to capture an entire life on video found that if the life were recorded on low web video, at 50 kbits/sec, the total space required would be 15TB. Even with DVD quality recording, t 5 Mbits/sec, the total storage would be 1,500TB. Clearly, an entire life can now be captured and stored on existing media.

In addition, while academic social scientists are increasingly using these cyber tools to combine data from a variety of sources – including text, video images, wireless network embedded devices and increasingly sophisticated phones, RFID's,[13] sensor webs, smart dust and cognitive neuron-imaging records, the same is also true of the private sector.

> Workers in warehouses across Britain are being "electronically tagged" by being asked to wear small computers to cut costs and increase the efficient delivery of goods and food to supermarkets, a report revealed yesterday. . . Under the system workers are

[10] Cyber infrastructure is a term coined by NSF to describe new research environments which exploit the newly available computing tools to the highest available level. These include computational engines (supercomputers, clusters, workstations – capability and capacity), mass storage (disk drives, tapes, . . .) and persistence networking (including optical, wireless), digital libraries/data bases, sensors/effectors, software (operating systems, middleware, domain specific tools/platforms for building applications), and services (education, training, consulting, user assistance). See Atkins et al. (2003). Revolutionizing Science and Engineering Through Cyberinfrastructure. Arlington, VA: NSF for more information. Available at http://www.nsf.gov/cise/sci/reports/atkins.pdf.

[11] http://www.nsf.gov/funding/pgm_summ.jsp?pims_id = 13553&org = CISE&from = fund.

[12] SBE/CISE workshop, March 15-16 2005, http://vis.sdsc.edu/sbe/About

[13] Radio frequency identification, or RFID, is a generic term for technologies that use radio waves to automatically identify people or objects. There are several methods of identification, but the most common is to store a serial number that identifies a person or object, and perhaps other information, on a microchip that is attached to an antenna (the chip and the antenna together are called an RFID transponder or an RFID tag). The antenna enables the chip to transmit the identification information to a reader. The reader converts the radio waves reflected back from the RFID tag into digital information that can then be passed on to computers that can make use of it. Source: http://www.rfidjournal.com/article/articleview/207

> asked to wear computers on their wrists, arms and fingers, and in some cases to put on a vest containing a computer which instructs them where to go to collect goods from warehouse shelves. The system also allows supermarkets direct access to the individual's computer so orders can be beamed from the store. The computer can also check on whether workers are taking unauthorised breaks and work out the shortest time a worker needs to complete a job (Hencke, *The Guardian* 2005).[14]

The capacity for this new technology to push forward the frontiers of social science research and answer important societal questions is clear. However, the progress will also put substantial pressure on statistical agencies to create and provide access to such data in order to keep pace with the private sector. Obvious new confidentiality challenges arise with these advances – such as protecting the identity of individual video images. The cartoon in Figures 1 and 2 is particularly illustrative.[15]

In addition to new data collection modalities, advances in cyber infrastructure also mean that much more administrative data can be stored and disseminated. As Pat Doyle often noted,[16] U.S. Census Bureau research has shown that the wide availability of certain kinds of personal information increases the chance of disclosure of confidential information – particularly when date of birth and geography are available. Yet, many states provide open access to administrative records that people can use to identify respondents. For example:

3.1. Birth Records

In most states, people who wish to obtain a birth certificate must demonstrate a need. However, some states (including California and Texas) are "open record" states. California birth records for 1905–1995 are available on the state web site and include the person's full name, birth date, sex, mother's last name, and county of birth. California's "nonidentifying births summary" database for 1996–1997 contains information on a person's county of birth, birth date, sex, race/ethnicity; mother's birth date, race/ethnicity, and state of birth; and father's birth date and race/ethnicity.

3.2. Marriage Records

Kentucky has a database on the web containing records for 1.1 million marriages (1973–2002). The database contains the name, age, race, residence, and number of prior marriages for the groom and the bride, as well as the date and county of the marriage and the marriage certificate number.

While marriage or birth records alone cannot be used to reidentify individuals on appropriately disclosure-proofed public-use microdata files, they can be used to enhance the information available from other sources, increasing the risk of disclosure.

Although there have been substantial advances in statistical disclosure protection techniques (see, for example, Winkler 2005, and some of the ideas put forward at a recent

[14] For the full report, see http://www.gmb.org.uk/shared_asp_files/uploadedfiles/95420EED-6333-4746-9BC0-432145FDD379_RegionalDistributionCentres.doc

[15] Thanks to Sang Kim and Chris Bratten for supplying the cartoon.

[16] The following three paragraphs are taken in their entirety from a working document authored by Pat Doyle, Julia Lane and Laura Zayatz (with permission from Laura).

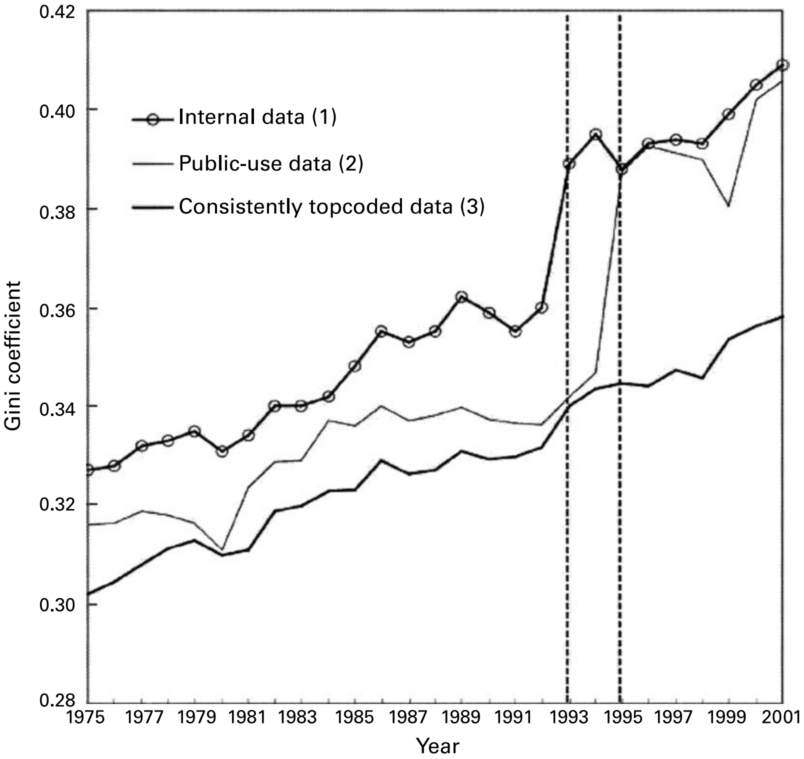

Fig. 1. Differences in earnings inequality trends with different data access

GOOD AFTERNOON MR. HENFENNEL! HEMORRHOID CREAM AISLE SIX!
BEEP!
TOO MUCH INFORMATION IN MY MEDICAL I.D. CHIP.
DON'T FORGET MOUTHWASH. I FIGURED THAT ONE OUT BY MYSELF. —
TOLES
10/6 ©2004 THE WASHINGTON POST

Fig. 2. Potential consequences of future data access

workshop organized by Dwork and Fienberg 2005) in response to some of these disclosure threats, little of this has been accompanied by a discussion of the effect on data quality, although Kaufman et al. (2005) do discuss the effect of masked procedures on data quality for a particular survey. This lack of attention is a major threat to high-quality empirical social science research, given the degradation in quality stemming from the use of current disclosure protection techniques and documented in the previous section.

4. Different Access Modalities

Public use microdata files are by no means the only access modality available to national statistical agencies. Others include research data centers, remote access facilities, and licensing.

Research data centers have been in use for as many as twenty years in some National Statistical Offices (NSO). In this modality, authorized researchers physically go to access data on a site controlled by the NSO, and are monitored by NSO employees. The computers within the Research Data Centers (RDCs) are not linked to the outside world; researchers do not have email or World Wide Web access from within RDCs. All analysis must be done within the RDC. Furthermore, there is typically an extensive review process to ensure that their work fits within the mandate of the NSO. As a consequence, researchers at the RDC may use confidential data only for the purpose for which the data are supplied; i.e., for their approved research project, and they may not remove confidential data from the RDC. There is also typically full disclosure review.

Although the RDCs have been effective in controlling identification risk particularly for data sets where a confidentialized microdata file is not possible, such as business data, they still require conditions of access to provide an adequate level of protection. The main criticism of DLs has been the lack of convenience to the researcher, including sometimes being forced to use unfamiliar data analysis software. They are also expensive for the NSO to manage compared with other options. A major concern is the length of the review process, the cost in terms of time and money, as well as the disparate effect caused by the distance some researchers have to travel to get to the RDC.

The key characteristic of remote access facilities is that researchers do not have to physically go to the NSO to work with the microdata.

There are two types of remote access:

(a) Buffered remote access. This approach permits the researcher to submit programs from a remote site, but does not permit the researcher to see the microdata. This is achieved either by sending the resultant output separately or by restricting the type of analysis that can be performed.
(b) Online remote access to the microdata with technical and legal protections against disclosure.

One of the oldest and best-known examples of buffered remote access is the Luxemburg Income Study, which permits research to submit batch programs, provides disclosure screening of output and returns output to users within 24 hours. This approach has led to underutilization of the resource, since the resultant delay in identifying coding errors led to too high a burden on researchers. Learning from this approach, Statistics Canada provides researchers with dummy microdata files so that they can test and debug their programs.

The Australian Bureau of Statistics also permits trial runs against small numbers of unidentifiable unit records to allow the identification of outliers, but only allows confidentialized microdata files to be accessed through the remote access facility. In all cases, output is checked before being returned to the user. Although these approaches have some appeal in terms of the perception that the data are being protected, they have substantial drawbacks. In order for research to be successful, it is necessary for researchers to be able to work directly with the microdata. This is particularly important in the case of outliers. In determining the factors contributing to economic growth, for example, it is critical to know whether a high income individual (or a high growth business) has been correctly identified as such, or whether there is a data entry error. In addition, the delays entailed by the layers of review before any output is seen places a high burden on the statistical agency and results in often unacceptable delays for decision-makers.

Recognizing these drawbacks, increasing numbers of statistical agencies are moving to online remote access systems. This approach uses modern computer science technology, together with researcher certification and screening, to replace the burdensome, costly and slow human intervention associated with buffered remote access.

The UK Office for National Statistics (ONS),[17] for example, instituted a full "remote laboratory" service in January 2004. Their approach is to use a thin client service, which means there is no data transfer at the user end. They have also centralized data management operations, which makes it much more efficient to work across different sites. Statistics Denmark[18] has found that remote access arrangements are now the dominant mode of access to microdata. Statistics Sweden's system for remote access to microdata (MONA[19]) provides users with secure access to databases at Statistics Sweden from almost any place with internet access. In this manner, Statistics Sweden has increased the accessibility of microdata for external users at the same time that it has increased security precisely because the client's computer functions like an input/output terminal. All application processing is done in the server.

In all cases, there are substantial advantages to the agency. New versions of the data can be made available without needing to produce disks or tapes for redistribution. The agency can create an easy to use front end.

Statistics Netherlands has gone even further in terms of its remote access. It has begun a pilot project, called the OnSite@Home facility[20] which makes use of biometric identification – the researcher's fingerprint – to ensure that the researcher who is trying to connect to the facility is indeed the person he or she claims to be.

Licensing is used by a variety of agencies. The approach involves the agency entering into a signed agreement with an external researcher that permits them to access semi-anonymized data files using a defined set of protocols at their home institution. The license

[17] Felix Ritchie "Access to Business Microdata in the United Kingdom" paper presented at the Joint UNECE/Eurostat work session on statistical data confidentiality (Geneva, Switzerland, 9-11 November 2005)

[18] Lars Borchsenius "New Developments in the Danish System for Access to Microdata" paper presented at the Joint UNECE/Eurostat work session on statistical data confidentiality (Geneva, Switzerland, 9-11 November 2005)

[19] Lars-Johan Söderberg. MONA – Microdata On-Line Access at Statistics Sweden, paper presented at the Joint UNECE/Eurostat work session on statistical data confidentiality (Geneva, Switzerland, 9-11 November 2005)

[20] Anco Hundepool and Paul-Peter de Wolf "OnSite@Home: Remote Access at Statistics Netherlands", paper presented at the Joint UNECE/Eurostat work session on statistical data confidentiality (Geneva, Switzerland, 9-11 November 2005)

typically includes a Data Security Plan that defines location, security arrangements and access protocols; confidentiality pledges; institutional concurrence, disclosure review, onsite security inspections and terms for termination.

5. An Economic Framework

It is clear that there is a plethora of access modalities: the issue for statistical agencies is finding the optimal modality (or combination of modalities). Trottini (2001) has developed a decision-theoretic framework to guide statistical agencies in their data release decisions. However, the decisions could be put in an economic framework as well. In such a framework, the data custodian is charged with maximizing data utility subject to both cost and disclosure constraints.[21] Each of these is discussed in more detail below.

There is a full discussion of the utility of microdata in Lane (2003a,b). Assume that the mission of each statistical agency is to maximize the utility to society, conditional on keeping disclosure risk at a predefined level.

Define U as data utility, the value to society of microdata access. This utility depends on data quality, researcher quality, and the number of times the data are accessed. Let $Q =$ Data quality, $R =$ Researcher quality, and $N =$ number of times the data are accessed.

Then we have $U = u(Q, R, N)$

Data quality depends on the portfolio of access modalities available to the research community. If $M_i =$ modality i, then we can write $Q(M_i)$. R and N are both determined by the access costs, A, imposed by the access modality, and we can therefore write R and N as functions of A: $R(A_i)$ and $N(A_i)$.

The expected costs to society of microdata access can be defined as the harm to individuals or organizations should disclosure occur, H, times the probability of disclosure, D, plus the monetary cost of providing access, C. The probability of disclosure is typically set at a "target" level: since most agencies are charged with using reasonable means to protect data, this implicitly means setting reidentification risk to some fixed number.

Thus, the expected social cost, S, can be written as

$$S = HD + C$$

The factors contributing to the target risk of disclosure D^* can be written as

$$D^* = d(E, I, Z, M_i)$$

where

E is the existence and accessibility of other data sources that can be used for reidentification. The relationship between this and reidentification is affected by technology, T, and can be written $E(T)$.

[21] This line of reasoning is heavily influenced by discussions with Pat Doyle and John Abowd as well as the work of Mark Elliott 2001.

I is the existence of malevolent interlopers. This relationship is affected by technology, legal penalties, L, and the characteristics of the population, X and can be written $I\,(T, L, X)$.

Z is researcher error. This is affected by technology, legal penalties, training and adoptable protocols, P, and can be written $Z(T{,}L, P)$.

M, as before, is the set of access modalities.

Harm, H, is also likely to be a function of population characteristics, and can be written $H(X)$.

Finally, the monetary cost constraint is

$$C = p_t\, T + \Sigma_{Mi} p_{Ai} M_i$$

where p_i reflects the price of providing a certain level of protection.

The constrained optimization is then to maximize utility subject to the constraint

$$S - C - HD^* \leq 0$$

or maximize the associated Lagrangian

$$L = U - \lambda\,(H\, d(E, I, Z,\, M_i) + p_t T + \Sigma_{Mi} p_{Ai} M_i - S)$$

In general, maximization requires that the marginal benefits with respect to each variable are set equal to the marginal costs. This, in turn, means that the statistical agency needs to be able to quantify the relative marginal value of each of the key input variables, which is no trivial task. And even this outline is relatively simplistic, since there are many potential measures of data usefulness: a fully comprehensive approach might well optimize over a multivariate utility space. The following section offers some suggestions towards this goal.

5. Using the Framework to Shape a Research Agenda

This framework, despite the somewhat cumbersome notation, serves the important function of identifying key focus areas for confidentiality research, namely:

5.1. Measuring the Value of High-Quality Data

The examples given in Section 2 highlight the serious consequences of data alteration. A natural extension of these examples would be to develop a methodology to value high-quality data. Although a natural lower bound might be the amount of money spent by nations on statistical agencies – for example, over $US 2 billion in the United States;[22] $A 338 million in Australia,[23] and SK 911 million in Sweden[24] – more scientific approaches exist. In particular, government and academic economists routinely put dollar values on human lives as a result of government regulation (see, for example, Lutter, Morrall, and Viscusi 1999 or Viscusi and Aldy, 2003). Similarly, extensive research has been done by Don Coursey to elicit the value of public goods, albeit in the context of environmental

[22] Ed Spar, "Federal Statistics in the 2007 budget", http://members.aol.com/copafs/AAAS2007.htm, Council of Professional Associations on Federal Statistics.

[23] http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1001.02005-06?OpenDocument

[24] http://www.scb.se/templates/Listning1____44031.asp: of which SK441 million is appropriated funds.

quality. A similar endeavor for statistical agencies would both serve to provide metrics of the effect of data distortion on public policy and highlight the value of data collection to the broader public.

5.2. *Developing Metrics of Data Quality **Q***

The work of Domingo Ferrer and Torra (2001a,b) and Duncan et al. (2001) which attempted to quantify information loss at the same time as measuring disclosure risk began to outline an approach that should be further advanced. Shlomo (2005) proposed a series of measures for frequency tables such as:

Distance metrics to measure distortions to distribution and expected totals,

Nonparametric statistical testing for same location, scale and shape of empirical distributions,

Effect on statistical inference, such as: Variance of cell size, Chi-Square measures of association, Pearson and log-likelihood ratio testing for log-linear modeling, and "Between" variation of target variables as expressed by R2.

However, similar metrics for microdata have not been developed. The two examples provided in Section 2 are illustrative of the issues in that measures of inequality require knowledge of the entire distribution; accurate measurement of key coefficients requires knowledge of the relationship among variables. This point has also been made by Winkler (2005b, c, and d), who notes the importance of developing measures that reflect the specific analytic use of the files.

Multiple approaches could be taken to determine these uses. One might be to undertake a literature review that summarized the main uses for major public use data sets; another to survey key federal and academic users.

5.3. *Quantifying the Effect of the Cost of Access **A** on Usage **N** and Researcher Quality **R***

The work by Dunne (2001) and Seastrom (2001) outlined some of the key issues associated with imposing high costs on researcher access. In the NSF award that served as one of the forces initiating the Longitudinal Employer-Household Dynamics program, Abowd, Haltiwanger, and Lane (1998) pointed out that for more than two decades, public policy around the world was influenced by analysis of public-use American microdata samples. However, the increasing availability of administrative data, as well as data from other countries, combined with the cost (including the cost of time) of accessing U.S. federal data, now means that many of the best researchers in the country, and in the world, have found alternative data sources for their empirical analysis.

Quantifying the effect of the cost of access, and using this as a basis for informed decision-making, would clearly be difficult. However, one possible approach would be to survey ten years of the relevant academic and federal literature and document how often federal data are used as a basis for analysis, relative to other sources, as well as identify any trends. Similarly, a survey of top federal and academic researchers would help identify the relationship between access and use.

5.4. Measuring Harm ***H***

Madsen (2003) outlined many of the key philosophical issues in an NSF workshop held in 2003.[25] He identified a key privacy paradox as follows:

> The "privacy paradox" occurs when data managers interpret the right to privacy as a near absolute ethical standard. Such an understanding of the nature of the right to privacy leads to an extreme understanding of the nature of the responsibility of confidentiality with newer and more restrictive controls on data access. More privacy in the research context paradoxically results in less social benefit, rather than in more (p. 3).

Researchers such as Singer (2001) and Greenia et al. (2001) have attempted to quantify harm, but an extensive research agenda remains, as first outlined by Lambert (1993). Both Greenia and Singer have since noted that the research agenda has also substantially changed since the events of September 11, 2001, both because government data collection activities have increased and because public perception of the harm associated with such collection is likely to have changed.

5.5. Quantifying the Relationship Between Other Data Sources ***E*** *and Disclosure* ***D***

Both Winkler (2003a and b, 2004 b, c, and d and 2005a) and Domingo-Ferrer and Torra (2001a,b, 2003) have outlined extensive research agendas.

5.6. Modeling Malevolent Behavior ***I*** *and Researcher Error* ***Z***

A recent NSF workshop on cyber infrastructure and the social sciences included, as one theme, the importance of using social science to understand and model malevolent behavior.[26] As was pointed out, the importance of this goes far beyond the federal statistical community, since such behavior affects a wide variety of realms – ranging from financial and personal harm (data and money, identity theft) to cyber-terrorism, "phishing" and "pharming," denial of service attacks, hacktivism, hate crimes, gambling and pornography. The summary report (see Berman and Brady 2005), noted that in this area:

> Social scientists can be especially helpful in developing an understanding of the motivations and capacities of those who might engage in malevolent behavior, in designing institutions and procedures that deter malevolent behavior and that produce trustworthy cyber infrastructure.

Indeed, there is a group of researchers – such as Joan Feigenbaum and Deb Agarwal – that has established a strong knowledge base in trust management issues and collaborative computing environments. Salvatore Stolfo and Roy Maxion have similarly extensive

[25] For a summary of the workshop, see Lane (2003a,b).

[26] Stephen Fienberg was the social science coordinator of this session; Shankar Shastry was the computer science coordinator.

research agendas to detect data mining based intrusion and to develop behavior based computer security models.[27, 20]

Hence, a sensible research agenda for the statistical community might well be to join forces with researchers to better model malevolent behavior, and develop sensible deterrents. The corollary would be to combine resources with other federal and private institutions that have common concerns.

*5.7. Investigating Alternative Technological Approaches **T** to Providing New Access Modalities **M***

Protecting databases against intruders has a long history in computer science (a classic article is Dobkin, Jones, and Lipton 1979). Computer scientists themselves are interested in protection of the confidentiality of the data on which they do research (for example, the Abilene Observatory supports the collection and dissemination of network data, such as IP addresses).[28] Cyber infrastructure advances have certainly served to expand the set of access modalities, particularly with respect to remote access. The cyber trust initiative at NSF has created an entire research community that focuses on creating network computers that are more predictable and less vulnerable to attack and abuse, that is developed, configured, operated and evaluated by a well-trained workforce, and that educates the public in the secure and ethical operation of such computers. The Department of Defense has developed different levels of web-based access ranging from unclassified (nipr-net) to secret (sipr-net) to top-secret (jwics-net)[29] using off the shelf technology. Similarly, the PORTIA project focuses on both the technical challenges of handling sensitive data and the policy and legal issues facing data subjects, data owners and data users. Finally, the recent NSF SBE/CISE workshop on cyber infrastructure[30] outlined a combined computer and social science research agenda for different approaches to access.

In addition, several agencies have preexisting institutional structures that could be used to expand the number and types of access modalities: such as the U.S. Census Bureau's Research Data Centers and the data enclave at NCHS. Similarly, the supercomputer centers funded by the National Science Foundation could be deployed to provide a portal for information about technical and nontechnical advances in confidentiality research, provide training about confidentiality procedures for researchers and institutional review boards and provide computational facilities to develop both technical and nontechnical solutions to confidentiality problems. Finally, the European Union is also making a substantial investment in a centralized location for social science data, and in the associated confidentiality issues, as part of its VIIth Framework.

6. Summary

Economists should act to promote the view that the federal statistical agencies, and other data custodians, should be as concerned about providing data for their customers and about

[27] See Project IDS http://www1.cs.columbia.edu/ids/index101503.html
[28] http://abilene.internet2.edu/observatory/
[29] I am grateful to Carl Landwehr for making me aware of this.
[30] SBE/CISE workshop, Match 15-16 2005, http://vis.sdsc.edu/sbe/About.

promoting use of their data as they are about protecting their respondents and ensuring the security of confidential information. The activities needed to avoid what some have called a pending "train wreck" between respondents, data custodians and data users involve technological advances, legal strategies, policy enhancements (related to both privacy and disclosure avoidance both in the context of survey and census data and in the context of administrative data), interagency coordination, new disclosure avoidance techniques, and privacy research.

This article has attempted to formalize a number of the issues and ideas that have circulated in disparate arenas. It began by noting that the study of confidentiality remains quite piecemeal in nature, without an overarching framework to provide a context. It highlighted the particular problems posed by a pursuit of confidentiality protection that did not pay attention to the main aim of providing data access, namely data utility, arguing that this could distort information and potentially lead to incorrect decisions. It outlined a standard economic approach to thinking about the optimization problem, provided a brief list of new initiatives and outlined a possible research agenda for optimizing access to microdata.

8. References

Abowd, J. and Lane, J. (2003). The Economics of Data Confidentiality. Mimeo, Committee on National Statistics. www7.nationalacademies.org/cnstat/Abowd_Lane.pdf.

Abowd J., Haltiwanger, J., and Lane, J., (1998). Dynamic Employer-Household Data and the Social Data Infrastructure. National Science Foundation, SES-9978093, September 28, 1999- September 27, 2003.

Berman, F. and Brady, H. (2005). Final Report: NSF SBE-CISE Workshop on Cyber Infrastructure and the Social Sciences, May. Available at www.sdsc.edu/sbe/.

Bernstein, J. and Mishel, L. (1997). Has Earnings Inequality Stopped Growing? Monthly Labor Review, December, 3–16.

Burkhauser, R., Butler, J., Feng, S., and Houtenville, A. (2004). Long Term Trends in Earnings Inequality: What the CPS Can Tell Us. Economics Letters, 82, 295–299.

Chay, K. and Powell, J. (2001). Semiparametric Censored Regression Models. Journal of Economic Perspectives, 15, 29–42.

Dobkin, D., Jones, A., and Lipton, R. (1979). Secure Databases: Protection Against User Influence. ACM Transactions on Database Systems (TODS), 4, 97–106.

Domingo-Ferrer, J. and Torra, V. (2001a). Disclosure Control Methods and Information Loss for Microdata. In P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (eds.). Confidentiality, Disclosure and Data Access. North-Holland: Amsterdam, 91–110.

Domingo-Ferrer, J. and Torra, V. (2001b). A Quantitative Comparison of Disclosure Control Methods for Microdata. In P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (eds.), Confidentiality, Disclosure and Data Access. North-Holland: Amsterdam, 111–133.

Domingo-Ferrer, J. and Torra V. (2003). Advanced Record Linkage for Disclosure Risk Assessment. Mimeo presented at NSF Workshop on Confidentiality.

Doyle, P., Lane, J., Zayatz, L., and Theeuwes J. (eds) (2001). Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, North Holland: Amsterdam.

Duncan, G. (2004). Exploring the Tension between Privacy and the Social Benefits of Govermental Databases. Mimeo, Carnegie Mellon University.

Duncan, G., Fienberg, S.E., Krishnan, R., Padman, R., and Roehrig, S.F. (2001). Disclosure Limitation Methods and Information Loss for Tabular Data. In P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (eds.). Confidentiality, Disclosure and Data Access. North-Holland: Amsterdam, 135–166.

Duncan, G.T., Keller-McNulty, S., and Stokes, S.L. (2003). Disclosure Risk vs. Data Utility: The R-U Confidentiality Map. Technical Report 2003-6. Heinz School of Public Policy and Management, Carnegie Mellon University.

Dunne, T. (2001). Issues in the Establishment and Management of Secure Research Sites. In Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes (eds). North Holland: Amsterdam.

Fienberg, S.E. and Slavkovic, A.B. (2004). Making the Release of Confidential Data from Multi-way Tables Count, Chance, 17(3), 5–10.

Greenia, N., Jensen, J.B., and Lane, J. (2001). Business Perceptions of Confidentiality. In Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes (eds). North Holland: Amsterdam.

Haworth, M., Bergdahl, M., Booleman, M., Jones, T., and Madaleno, M. (2001). LEG Chapter on Quality Framework. Proceedings of Q2001, Stockholm, Sweden, May. CD-ROM.

Hencke, D. (2005). Firms Tag Workers to Improve Efficiency. The Guardian, Tuesday June 7.

Kaufman, S., Seastrom, M., and Roey, S. (2005). Do Disclosure Controls to Protect Confidentiality Degrade the Quality of the Data? Proceedings of the American Statistical Association, Section on Survey Research.

Lambert, D. (1993). Measures of Disclosure Risk and Harm. Journal of Official Statistics, 9, 313–331.

Lane, J. (2003a). The Uses of Microdata. Keynote Speech to Conference of European Statisticians, Geneva, Switzerland. http://www.unece.org/stats/documents/ces/2003/crp. 2.e.pdf.

Lane, J. (2003b). Key Issues in Confidentiality Research: Results of an NSF Workshop, May. http://www.nsf.gov/sbe/ses/mms/nsfworkshop_summary1.pdf.

Lerman, R. (1997). Reassessing Trends in Earnings Inequality. Monthly Labor Review, December, 17–25.

Lutter, R., Morrall, J.F.III, and Viscusi, W.K. (1999). The Cost-per-Life-Saved Cutoff for Safety-Enhancing Regulations. Oxford University Press, Economic Inquiry, 37, 599–608.

Madsen, P. (2003). The Ethics of Confidentiality: The Tension between Confidentiality and the Integrity of Data Analysis in Social Science Research. Mimeo, Carnegie Mellon University.

Mishel, L. and Jared B. (1997). Has Wage Inequality Stopped Growing? Monthly Labor Review December, 3–16.

Seastrom, M. (2001). Licensing. In Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (eds). New York: Elsevier, 341–370.

Shlomo, N. (2005). Information Loss Measures for Frequency Tables. Mimeo, Southampton Statistical Sciences Research Institute.

Singer, E. (2001). Public Perceptions of Confidentiality and Attitudes Toward Data Sharing by Federal Agencies. In Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (eds). New York: Elsevier, 341-370.

Smith, J. (1991). Data Confidentiality: A Researcher's Perspective. Proceedings of the American Statistical Association, Section on Social Statistics, 117–120.

Trottini, M. (2001). A Decision-Theoretic Approach to Data Disclosure Problems. Research in Official Statistics, 4, 7–22

U.S. Department of Labor (2002). Current Population Survey Design and Methodology, Technical Paper TP63RV. Washington DC.

Viscusi, W.K. and Aldy, J.E. (2003). The Value of a Statistical Life: A Critical Review of Market Estimates throughout the World. Journal of Risk and Uncertainty, Springer, 27, 5–76.

Winkler, W.E. (2003a). Methods for Evaluating and Creating Data Quality. Proceedings of the ICDT Workshop on Cooperative Information Systems, Sienna, Italy, January. Longer version in Information Systems (2004), 29, 531–550.

Winkler, W.E. (2003b). Data Cleaning Methods. Proceedings of the ACM Workshop on Data Cleaning, Record Linkage and Object Identification. Washington, DC, August.

Winkler, W.E. (2004b). Masking and Re-identification Methods for Public-Use Microdata: Overview and Research Problems. In Privacy in Statistical Databases, J. Domingo-Ferrer and V. Torra, (eds.). New York: Springer, 231-247. http://www.census.gov/srd/papers/pdf/rrs2004-06.pdf.

Winkler, W.E. (2004c). Record Linkage: Overview of Recent Developments and Applications. In Combining Data from Different Sources – Applications of Record Linkage Methodology and Estimation Using Administrative Data, S. Biffignandi (ed.). Rome: ISTAT.

Winkler, W. (2005). Microdata Confidentiality References. Mimeo, U.S. Census Bureau, 11 February.

Winkler, W.E. (2005a). Overview of Record Linkage and Current Research Directions. U.S. Bureau of the Census, Statistical Research Division Report at http://www.census.gov/srd/www/byyear.html.

Winkler, W.E. (2005b). Data Quality in Data Warehouses. In Encyclopedia of Data Warehousing and Data Mining, J. Wang, (ed.).

Winkler, W.E. (2005c). Methods and Analyses for Determining Quality, to appear.

Winkler, W.E. (2005d). Data Quality for Modeling, Analysis, and Data Mining.

Winkler, W.E. (2005e). Methods and Analyses for Determining Quality, 2nd Keynote address at the 2005 *ACM SIGMOD Workshop on Information Quality in Information Systems* (available under Post Workshop Material at http://iqis.irisa.fr/).

Winkler, W.E. (2005f). Modeling and Quality of Masked Microdata. Proceedings of the American Statistical Association, Section on Survey Research Methods.

Zayatz, L. (2005). Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update. Research Report Series (Statistics #2005-06), Statistical Research Division, U.S. Census Bureau, Washington, D.C.

Received October 2005
Revised February 2007