Towards a National Infrastructure for Community Statistics

TOOLS TO AVOID DISCLOSING INFORMATION ABOUT INDIVIDUALS IN PUBLIC USE MICRODATA FILES

Cynthia M. Taeuber
The Jacob France Center, University of Baltimore

A Discussion Paper Prepared for the
The Brookings Institution Metropolitan Policy Program

June 2006

URBAN MARKETS INITIATIVE
SUMMARY OF PUBLICATIONS*

2006

Fulfilling the Promise: Seven Steps to Successful Community-Based Information Strategies

The Affordability Index: A New Tool for Measuring the True Affordability of a Housing Choice

2005

Federal Statistics: Robust Information Tools for the Urban Investor

Market-Based Community Economic Development

Using Information Resources to Enhance Urban Markets

2004

Using Information to Drive Change: New Ways of Moving Markets

* Copies of these and Brookings metro program publications are available on the web site, www.brookings.edu/metro/umi, or by calling the program at (202) 797-6131.

ACKNOWLEDGMENTS

The author thanks John Abowd for his valuable comments and the supplemental materials he provided for this review of data confidentiality tools. Thanks also to William Winkler for his indispensable advice on the impact of advances in record linkages as well as his permission to include the bibliographies he compiled on microdata data confidentiality (Appendix A) and record linkages (Appendix B). The author also thanks Andrew Reamer, Laura Smith, Alan Karr, Jerry Reiter, Katherine Wallman, Susan Schechter, and Brian Harris-Kojetin for their help and comments.

This paper was produced with generous support from the Ford Foundation and the Rockefeller Foundation.

ABOUT THE AUTHOR

Cynthia M. Taeuber is the principal of CMTaeuber & Associates and a Research Associate at the Jacob France Institute at the University of Baltimore where she is responsible for managing and coordinating projects undertaken by the Institute in the area of community statistics and data produced by federal statistical agencies. With 30 years of experience at the U.S. Census Bureau, her work encompassed the development of statistics to address public policy issues and the transformation of demographic and economic data into knowledge for decision-makers. Ms. Taeuber is an acknowledged expert on decennial census data, the American Community Survey, the Local Employment Dynamics program, and the uses of administrative records in community information systems. Ms. Taeuber authored major books and publications on diverse demographic and economic topics including the older population, older workers, and women in the American economy. She holds a Master's degree in demography from Georgetown University and a Bachelor's degree in political science from the University of Texas.

Comments about this paper can be directed to Cynthia Taeuber at cmtaeuber@hughes.net.

ABOUT THE PUBLICATION SERIES

This publication is one in a series that investigate the design, development, and implementation of National Infrastructure for Community Statistics (NICS). To find out more information about NICS go to http://www.nicsweb.org.

METROPOLITAN POLICY PROGRAM

EXECUTIVE SUMMARY

Statistical agencies walk a fine line meeting their legal obligation to avoid disclosing information about an individual while still providing useful data for research on complex policy questions.

To meet the needs of data users (researchers and policy makers), statistical agencies have long provided public use microdata (PUMS) files for research that also have a low risk of re-identification of individuals. Two technological advances of recent years make it easier to create a "mosaic" of data sets that increase the chances of identifying an individual and make it more complex for statistical agencies to meet their statutory mandate to keep private information private:

- A great deal of information is available about individuals on the Internet.
- Sophisticated software has been developed that allows linkage of records that can identify a small percentage of individuals that make up traditional public use microdata (PUMS) files.

To protect confidentiality under these circumstances, national statistical agencies invest heavily in statistical techniques, software, and policies that safeguard the confidentiality of the data they release for public use.

This paper examines the risks and effectiveness of traditional techniques for protecting data confidentiality and privacy of individuals. It recommends that statistical agencies invest more than they have already to find alternatives to enhance data quality and lower the risks of re-identification of individuals. Many techniques developed for federal uses are applicable to community statistical systems—data sources at the state, regional, and local level.

TABLE OF CONTENTS

TOOLS TO AVOID DISCLOSING INFORMATION ABOUT INDIVIDUALS IN PUBLIC USE MICRODATA FILES

I. INTRODUCTION

Statistical agencies walk a fine line to meet the needs of two sets of customers: those who demand more detailed data to better understand complex policy questions, and those who demand that their responses to surveys or their use of public services be kept confidential.

To balance these two legitimate concerns, national statistical agencies invest heavily in tools and policies that safeguard the confidentiality of the data they release for public use.[1] State and local governments vary in their response. Some:

- Do not release the program records at all;
- Limit access to a small pool of trusted researchers;
- Release online protected data sets that use the same tools as the national statistical agencies (that is, limited data); or
- Release online unprotected data that reveal individual characteristics.

Public use microdata products are derived primarily from national sample surveys conducted by statistical agencies, or administrative records from state and local government programs. The national databases derived from surveys have the advantage of being comparable across areas. In contrast, administrative records in all but a few cases are generally idiosyncratic and specific to a state or local area. A strong advantage of administrative data is the lower cost to both data collectors and respondents. Administrative data are collected as part of the record-keeping process and therefore there is no additional burden to respondents as in a survey. In addition, they are low-cost sources of information. The Local Employment Dynamics Program, for example, costs approximately 2 cents per record to process compared with $50 or more for most surveys.

A drawback, however, is that new techniques for linking records may inadvertently identify individuals, and old techniques degrade data quality and inhibit research. With traditional statistical tools, disclosure can occur even when agencies are meticulously trying to release only data that are secure while allowing sufficient detail to respond to policy questions.[2] Respondent perceptions of data security also matter. Therefore, statistical agencies, which must by law protect confidentiality,

[1] For further details about national statistical policies and resources for confidentiality and data access information, see Confidentiality and Data Access Committee (CDAC) of the Federal Committee on Statistical Methodology, Office of Management and Budget: www.fcsm.gov/committees/cdac/resources.html.

[2] Statistical techniques that protect the confidentiality of the information provided by individuals or businesses can limit disclosure. W.E. Winkler, "Views on the Production and Use of Confidential Microdata." U.S. Census Bureau Research Report no. RR97/01 (Washington: U.S. Census Bureau, 1997); and W.E. Winkler, "Producing Public-Use Files That Are Analytically Valid and Confidential." U.S. Census Bureau Research Report no. RR98/02 (Washington: U.S. Census Bureau,1998).

and which also need high response rates for accuracy, pay attention to respondent concerns, both real and perceived, and work to allay any concerns to encourage cooperation.

The question then, for both surveys and administrative records, is how to release detailed and high-quality data from all data sources while protecting the identity of individuals, both businesses and residents, and the information they provide. Research on data security has accelerated, including research on how to limit or prevent disclosure of personal information, ensure valid analytic properties, and measure the risk of disclosure and the harm it does. Research has also expanded on methods to limit disclosure in microdata records and methods of analyzing data that have been disclosure proofed.[3]

The purposes of this paper are to:

1. Offer guiding principles for disclosure limitation tools when using Public Use Microdata Samples (PUMS) files from the U.S. Census Bureau;

2. Describe selected tools and methods of protecting the confidentiality of microdata and valid analytic properties; and

3. Assess the relevance of these tools for the National Infrastructure for Community Statistics (NICS), a proposed nationwide web-based utility that facilitates access by public and private decision-makers to detailed, current community-level statistics from thousands of local, state, federal, and commercial data sources.

[3] John M. Abowd and Simon D. Woodcock, "Disclosure Limitation in Longitudinal Linked Data." In Pat Doyle, et. al, eds., Confidentiality, Disclosure, and Data Access (Amsterdam: North Holland, 2001). Also see John M. Abowd and Simon Woodcock, "Multiple-Imputing Confidential Characteristics and File Links in Longitudinal Data." In J. Domingo-Ferrer and V. Torra, eds., Privacy in Statistical Databases (New York: Springer-Verlag, 2004), pp. 290-297.

II. BACKGROUND

A. The Conflicting Needs to Safeguard Individual Information and Provide Detailed Data

State and local governments have in recent years come to understand the potential value of converting program records to statistical files. However, this process is not a simple matter. Although less expensive than original data collection, program-record conversion involves real costs and carries the danger that, without safeguards, individual confidentiality could be violated. Policies and tools have been developed that begin to address this concern, as discussed in this paper.

Statistics are released as predefined summary (frequency count) tabulations or as public use microdata files, after applying a variety of disclosure avoidance methods. As Doyle and colleagues defined them, "*frequency count tables* count the number of respondents with specified characteristics....Tables of aggregate magnitude data are analogous to frequency count tables in that they are defined by cross-classification of categorical variables. However, the cells contain aggregate values over the corresponding respondents, of some quantity of interest."[4] Such tabulations are purposefully limited and are not sufficiently detailed for complex analyses.

"*Microdata files* consist of individual records that contain values of variables for a single person, a business establishment, or another individual unit. Public use microdata files are released to the public for research and analytical purposes after being subjected to procedures that limit the risk of disclosure."

Microdata files may be from one administrative source or linked together from different, integrated sources. Federal statistical agencies remove identifying information, incorporate precautions to ensure confidentiality, and subject all microdata files to strict disclosure reviews before they are released to the public, paying special attention to integrated files.

Complex policy questions call for detailed data sets. Abowd and Lane point out that most researchers prefer microdata so they can design cross-tabulations and separate demographic, economic, environmental, and spatial interactions.[5] In addition, researchers can readily replicate others' work, and they can perform multivariate analyses to isolate the marginal

[4]Pat Doyle and others, eds. Confidentiality, Disclosure, and Data Access (Amsterdam: North Holland, 2001), p. 4.

[5] Ronald Rindfuss, "Confidentiality Promises and Data Availability," IHDP Update: Newsletter of the International Human Dimensions Programme on Global Environmental Change (February 2002).

impact of key variables, controlling for other factors, rather than reaching more limited, and possibly misleading, conclusions on the basis of averages.[6]

And yet, as noted, avoiding the disclosure of personal information has long been a barrier to the release of data that are as detailed as researchers need to respond to complicated questions.[7]

B. Two Sides to Increased Computing Power

With increased computing power, much has become possible—with risks and advantages. The technology has:

- Increased the collection of data by the largely unregulated private sector;
- Provided online access to administrative records with unprotected individual characteristics, such as those derived from birth records, deaths, voter registrations, marriages, and drivers' licenses;
- Made it easier to link datasets by matching information about individuals, such as exact birth date and detailed geography of residence or workplace, thus:
  - Improving the ability of researchers to address complex, multivariate public policy questions; but also
  - Increasing the risk that individual people or businesses can be identified in a data set.
- Made it possible to display microdata on maps, including those online.

C. Response of Statistical Agencies

Statistical agencies balance the risk of disclosure with public policy benefits.[8] Although advances have been made in avoiding disclosure, they may not go far enough for public use microdata files. The growth in the number of administrative files, along with enhanced computer power, have increased the probability of identifying individuals in the public use microdata files. Federal agencies now are concerned they will no longer be able to provide public use microdata files from their surveys and programs to researchers using conventional tools.

---

[6] John M. Abowd and Julia I. Lane, "The Economics of Data Confidentiality" (Washington: Committee on National Statistics, National Academies of Science, October 2003), available at: www7.nationalacademies.org/cnstat/Abowd_Lane.pdf (May 2006).

[7] George T. Duncan and others, "Disclosure Limitation Methods and Information Loss for Tabular Data." In Pat Doyle and others, eds., Confidentiality, Disclosure, and Data Access (Amsterdam: North Holland, 2001).

[8] There are various federal laws and regulations that govern confidentiality rules and policies. For example, the Census Bureau collects its survey and census data under Title 13 of the U.S. Code and the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA). Both require the protection of the information collected and allow only statistical uses of the data. The goal is to release high-quality data without violating the promise of confidentiality of the information. See www.census.gov/privacy/files/data_protection/002775.html.

Traditionally, the data have been protected through laws, statistical techniques, and statistical policies, including rules for access to data. Traditional techniques to protect confidentiality include releasing data at higher geographic aggregations only, adding "noise" to the data, blanking some values, swapping data across households, and bottom- and top-coding characteristics such as income. The efforts, however, have some researchers complaining that they cannot properly analyze the data in ways useful to policymakers.[9] Federal statistical agencies, on the other hand, are concerned that the rules are not sufficient enough, given the proliferation of administrative records. To meet the need for access to microdata, statistical agencies have invested in research to develop more sophisticated approaches and have developed new access rules including licensing, remote access, and access through secure remote sites. Statistical agencies know they must be sensitive to perceptions about that access as well as to the legalities and technicalities of access.

[9] Doyle and others, Confidentiality, Disclosure, and Data Access, p. ix.

III. GUIDING PRINCIPLES FOR LIMITING DISCLOSURE IN PUBLIC USE DATA FILES

Understanding that there is a growing risk to confidentiality in public use data files, NICS needs to define its guiding principles for disclosure-limitation tools. Several researchers have begun to develop such principles (see Appendix A for a bibliography of recent research), including the following by Stephen Fienberg:[10]

- Inferences should be the same as those drawn from the original data;
- Researchers should be able to reverse disclosure protection tools for inferences about parameters in statistical models but not about individual identifiers;
- Multivariate analyses need sufficient variables;
- Researchers need enough summary information to examine outliers; and
- Researchers need enough information to judge the goodness of fit of models.

The Census Bureau's expert on record linkage and data confidentiality, William Winklernotes that the final data set for public use should resemble the original data set while retaining privacy. [11]

The difficulty, according to Winkler, is that none of the principles have yet to be demonstrated, if we understand inferences as referring to nontrivial analytic properties. By "analytic properties," Winkler means that if the original file with no confidentiality edits allows a type of analysis such as regression, then the masked data should allow the same analysis and reach nearly the same results.[12] Below we discuss two confidentiality methods, masked and synthetic data. Both are able to approximately reproduce a total, a mean, and a correlation from the data, the basic requirements of a data set that is acceptable for researchers. A problem with simply masking data, however, is that the more analytic properties and the more variables a public use microdata file has, the easier it is to re-identify. Winkler argues that the masking methods that are easiest to implement, such as additive noise, blanking values, swapping values across records, and recoding, "seldom, if ever, have been justified in terms of preserving one or two analytic properties or in preventing re-identification."[13] As such, below we provide a detailed look at the alternative – synthetic data.

[10] S. Fienberg, "Allowing Access to Confidential Data: Some Recent Experiences and Statistical Approaches." Presentation at Stockholm Workshop on Microdata Access, August 21, 2003.
[11] William E. Winkler, "Modeling and Quality of Masked Microdata." U.S. Census Bureau, Statistical Research Division, 2006, p. 1): available at: www.census.gov/srd/www/abstract/rrs2006-01.html, (May 2006).
[12] Personal correspondence with William E. Winkler, U.S. Census Bureau, September 2005.
[13] Winkler, "Modeling and Quality of Masked Microdata," p. 4.

IV. TOOLS TO LIMIT DISCLOSURE WHEN USING PUBLIC USE MICRODATA

There are currently three categories of confidentiality tools:

1. Restricted access under the watchful eye of the data owner;
2. Introducing uncertainty, including top coding, limiting geographic detail, suppressing information, and adding random noise, sometimes in several stages; and
3. Altered data, such as data swapping, micro-aggregation, and synthetic data models.

Alan Karr, the director of the National Institute for Statistical Sciences, refers to the second category as "the truth, but not the whole truth," and the third category as "the approximate truth but not the truth."[14] The categories are discussed in more detail below.

A. Restricted Access

Of the three options, access to the full source file is of greatest benefit to researchers. As we discuss in more depth below, approaches have been developed for this method including:

- A web-based tool for swapping data;
- Remote submission of research programs to the holder of the source file, with results being returned after examination to prevent disclosure of confidential data;
- Online query systems;
- Licensing procedures for selected data users to have access to confidential data; and
- Secure research data centers (RDCs) where researchers can work directly with the full source file.

Another potential tool is to develop technology to enforce accountability for protecting data confidentiality. Proposals to develop such technology have been suggested for the federal government's Data Reference Model project.[15]

1. NISS WebSwap: A tool to limit disclosure in microdata files

The National Institute of Statistical Sciences (NISS) hosts a web service, the NISS WebSwap, to protect individually identifiable data by swapping characteristics among records. As discussed below, data swapping is a traditional tool for limiting disclosure of information in confidential microdata files by introducing uncertainty into data. An individual record is altered by

[14] Alan Karr, "Data Swapping and Other Confidentiality Tools." Remarks at the National Infrastructure for Community Statistics (NICS), CoP Research Symposium on "The Emerging Tool Kit for NICS" (Washington: Brookings Institution, June 30, 2005).

[15] Stewart Baker and Jeff Jonas, "Appendix F: Technology Challenges for the Near Future." In Creating a Trusted Network for Homeland Security: Second Report of the Markle Foundation Task Force, a publication of the Practices of Dynamic Knowledge Repository-Semantic Web Services project (2003): available online at web-services.gov/lpBin22/lpext.dll/Folder17/Infobase6/1/50c/688/6f9?fn=main-j.htm&f=templates&2.0 (May 2006).

switching the value of characteristics across randomly chosen pairs of records in a small proportion of the original records.

The NISS web service uploads a user's original data file and then downloads an altered file with the swapped records.[16] The user specifies the fraction of records to be swapped (usually 1 to 10 percent) and constraints, if any, on the unswapped attributes. The WebSwap is currently a prototype and not production software. In addition, the service is not currently fully secure and cannot be used on truly confidential data.

2. Remote submission of programs for offline access to data

Remote offline access allows researchers access to the original, more detailed confidential data than is available in a public use file. User interface resides on a computer outside firewalls, and it contains no data. The interface accepts software programs such as SAS, and it determines that the program works as it should. The interface reads the request to another computer behind the firewall to execute the data request. The internal machine runs an automated disclosure review, and if the results pass, it sends them to the machine outside the firewall. The researcher receives results from the external machine.

Although seemingly a practical solution to confidentiality issues, this approach is not the first choice of most researchers because it is not interactive. It is more difficult for them to get a feel for the data and they can do less experimenting with tabulations to gain insight into the questions they are researching. It is troublesome for them to ask technical questions about the data, and the data owner receives less feedback about the limitations of the data set. The data set is then less likely to be improved than when there is direct interaction between data users and data owners.

3. Online query systems

DataFerrett (http://dataferrett.census.gov/) is a data mining and extraction tool hosted collaboratively by the U.S. Census Bureau and the Centers for Disease Control. It allows data users to select a data basket of variables and recode those variables as needed. With DataFerrett, a data user can develop and customize tables from the original full data file. DataFerrett allows data users to locate and retrieve data without charge, regardless of where the data reside. DataFerrett allows data providers to share their data more easily and manage their own online data. If they provide the funding, local areas can load their data onto DataFerrett for public access as well.

A second online query system, developed by NISS, is the National Agricultural Statistics Service (NASS) System of Geographical Aggregation. The system disseminates survey data about farms down to the county level. Originally, the data for more than one-half of the counties could not

[16] Ashish Sanil, Shanti Gomamtam, and Alan F. Karr, "NISS WebSwap: A Web Service for Data Swapping," Journal of Statistical Software 8 (7) (2003): 1-12, available at: www.niss.org/dg/TR/nisswebswap200206.pdf. Other references on this topic from NISS are available at www.niss.org/dg/technicalreports.html and www.niss.org/dgii/techreports.html.

be disclosed. As a result, NISS developed a web-based query system that merged undisclosed county data with neighboring counties in the same state.[17]

A third tool is a "table server" developed by NISS to disseminate marginal subtables of a large contingency table. The table server allows dynamic assessment of disclosure risk, in light of previous queries, "which also allows data to be probed most deeply in regions of user community interest."[18] One issue with this system is that it does not provide access to microdata files. A second issue is that "a table cannot be released if, together with previously released information, it would place the system in a state whose risk exceeds the threshold." [19] There are many other research issues associated with table servers that NISS is exploring.

4. Licensing selected users to have data access

One approach to data confidentiality is to secure an agreement between the researcher and the data owner to the terms of data use (including whether the data file can be matched with other files) and the penalties for misuse. For example, the National Center for Education Statistics (NCES) licenses selected data users with an "affidavit of nondisclosure," which allows them to hold confidential data. The agency is responsible for protecting the data and overseeing its use by the licensee, who is subject to random inspections and severe penalties for unauthorized disclosures.[20]

Licensing requires a high level of trust and carries a significant risk to the data owner. Regardless of who is at fault, a disclosure reflects more poorly on the data owner than on the data user. Massell and Zayatz note that those who sign license agreements do not always pay close attention to their responsibilities. They may not report, for example, changes in custodianship of the data.[21]

Licensing does not meet the NICS objective for broad access. Ideally, to ensure the data are being protected, the method requires recurring investigations by the data owner and review of reports before publication. This is an expensive undertaking that limits access without ensuring compliance. Further, who gets a license often depends on personal relationships of trust with state and local officials, further limiting access and raising charges that access depends on favoritism. Such is not the case with NCES and other federal statistical agencies that use licensing.

---

[17] Alan F. Karr and Ashish P. Sanil, "Web Systems that Disseminate Information But Protect Confidentiality" (Research Triangle Park, NC: National Institute of Statistical Sciences, n.d.): available at www.niss.org/dg/TR/karr-sanil-iass.pdf.
[18] Ibid, p. 1.
[19] Ibid, p. 3.
[20] U.S. Government Accountability Office, "Record Linkage and Privacy: Issues in Creating New Federal Research and Statistical Information." GAO-01-126SP (April 2001), p. 91.
[21] Paul B. Massell and Laura Zayatz, "Data Licensing Agreements at U.S. Government Agencies and Research Organizations." Proceedings of the International Conference on Establishment Surveys – II (Buffalo, NY: ICES 2000, June 2000). An example of licensing is the agreement that must be signed by anyone seeking to use data in the Nationwide Inpatient Sample. See sample at: www.ahcpr.gov/data/hcup/datause.htm.

Finally, if a large number of researchers have access to individual data under licensing, security of the data is reduced accordingly. Enforcement of penalties for misuse means there will be publicity about the breach, and bad publicity may be a disincentive to the data owner to pursue violators of the license.

5. Research data centers

The Census Bureau has established secure research data centers (RDCs), where researchers granted "special sworn status" have restricted access to parts of confidential micro records under a strict set of rules and time-consuming limitations.[22] The variables under study must be approved in advance; that is, the researcher does not have access to the entire data set. The research access must provide a benefit to the programs of the Census Bureau (see Title 13, Sec. 23, U.S.C.). The research must also have scientific merit, have a clear need for nonpublic data, be feasible with the confidential data (limited Census value-added), and the research output must pass a rigorous, multi-step disclosure review process by the Census Bureau and others as appropriate.

The RDCs are secure facilities in which computers are not linked to the outside world and, thus, do not provide e-mail or Internet access. All analysis must be done within the RDC. Researchers can use the confidential data for their approved project only, they cannot remove any confidential data from the site, and all data products (intermediate output and publications) must be submitted to the Census Bureau for review before the data can leave the RDC.

In short, RDCs give researchers access to the source data but only after an arduous process and under many limitations. In addition, there are only a limited number of RDC sites across the United States. Therefore, for most researchers, access entails travel expenses and substantial time away from home offices.

B. Introducing Uncertainty and Suppressing Information

All data sets are an approximation of the truth. They all have "nonsampling error," that is, errors from, for example, responses given by respondents or errors in processing the data. Surveys also have sampling error. It is therefore not unreasonable to consider methods of further altering data, that is, intentionally adding error to data elements to protect confidentiality, so long as the distributions and relationships to the unaltered data are maintained as closely as possible. Altering data has been the traditional basis for releasing public use microdata files.

For microdata, common procedures to avoid disclosure of personal information include introducing uncertainty (such as by swapping values among similar respondents and adding random noise) and suppressing information that directly or indirectly identifies an individual by rounding, top-

[22] Such researchers must undergo a security check, including fingerprinting. Researchers with a Special Sworn Status are subject to the same legal penalties as regular Census Bureau employees for disclosure of confidential information (that is, a fine of up to $250,000, imprisonment for up to five years, or both). See www.ces.census.gov/ces.php/rdc.

and bottom-coding, collapsing response categories, and providing information only for higher geographic levels. One can also average values, such as averaging the three highest values and inserting the average into each record. Rounding is often used in conjunction with other techniques.

As noted in a Government Accountability Office (GAO) report, certain variables are more risky than others, and therefore the choice of which variables to modify or eliminate can introduce potential limitations.[23] Because of the proliferation of online database searches about individuals, some variables (birth date, for example) carry higher risks of identifying individuals than others (income bracket, for example) because of the many other data sources that might contain the variable. For this reason, among others, longitudinal data files are particularly risky.

In addition, some methods of adding error may produce biased estimates, particularly with data swapping.[24] For example, data swapping may "distort joint distributions involving both swapped and unswapped attributes."[25] There is a large literature on methods of measuring and limiting bias.[26] The most appropriate method depends on the characteristics of the particular data set.

Abowd and colleagues have developed refined methods for adding noise to the data in the Quarterly Workforce Indicators (QWIs) of the Local Employment Dynamics (LED) program, as well as to generate protected geo-spatial matching of establishment and household data from that program. The resulting statistics are "fuzzed" data items. Some statistics are significantly distorted from the original statistics and flagged as such in an effort to preserve confidentiality. Other tools, including aggregation, limited suppression (where the number in a cell is very small), estimation procedures, and weighting processes are a further effort to protect the confidentiality of the information.

The degree of confidentiality that is provided by these methods has not been rigorously tested and is still an open question. Means are preserved and variation can be measured, for example.[27] Their research does indicate, however, that this method allows the release of masked LED statistics that are statistically valid, even when including the significantly distorted values. The "fuzzed data" were explicitly designed to permit accurate trend analyses, which is an example of how confidentiality protection measures can be tailored to enhance analytic validity. Cells that cover many employers in the LED, for example, have very little data distortion. In cells with relatively few employers, the values are distorted as little as necessary to maintain confidentiality. In either case, the quarter-to-quarter change is reliable. In addition, the statistical properties of the error associated with the estimate are reported to data users so that they can successfully use the information. Cells

[23] U.S. GAO, "Record Linkage and Privacy."
[24] Bias is defined as the deviation of the average survey value from the true population value.
[25] Sanil, Gomamtam, Karr, "NISS WebSwap," p. 1.
[26] Abowd and Woodcock, "Disclosure Limitation" and "Multiplying-Imputing Confidential Characteristics." See also Winkler, "Producing Public-Use Files."
[27] Personal communication with John Abowd, professor of industrial and labor relations, Cornell University, June 8, 2005.

that are "fuzzed" by a relatively large amount to protect the confidentiality of the underlying data are flagged in the data releases to warn data users of the significant distortion.

The details of the statistical methodology Abowd and his colleagues use to protect the data underlying the QWIs are to be published in the near future. The methods are complex, however, and the factors and direction of fuzz are confidential to protect the data releases.

For the purposes of NICS, these traditional methods of ensuring confidentiality are well known and therefore possible to apply to public use microdata. In reality, however, these traditional tools can be obstacles to the release of additional administrative records because:

- Specialized software is needed
- Sophisticated statisticians are needed to choose methods appropriate to the characteristics of the data set and how it will be used. Some of their tasks include:
  - Assessing the statistical characteristics of the data set such as whether the unmasked data have a multivariate normal distribution and whether that makes a difference in the appropriate choice of methods.
  - Assessing the analytic usefulness of the file, especially for small population groups or lower geographic levels, which may be compromised more so than larger aggregations such as the state or national data.
  - Using a method that maintains the correlation structure and the means.
  - Coordinating cell suppression techniques among tables, and often using secondary suppressions. A simpler alternative for establishment tabular data is to add noise to the original file before tabulating data and to add more noise to "more sensitive" cells.[28]
- Determining the risk of re-identification of the altered file.[29]
- The results yield files that are analytically useless and allow re-identification.[30] Also, the more data are masked by traditional methods, the less the final data set resembles the original data set, thus compromising analysis of the distribution and relationship of characteristics.
- Data linkage methods (see Appendix B for a bibliography on the topic) can often defeat the intent of traditional tools and lead to the identification of individuals.[31]

C. Synthetic Data Methods for Public Use That Are Valid for Analyses

Synthetic data are an alternative to traditional confidentiality tools. Synthetic data have essentially the same characteristics as the actual data. In synthetic data models, every record of actual data is replaced with synthetic data such that a record is no longer that of an individual.

---

[28] T. Evans, L. Zayatz, and J. Slanta, "Using Noise for Disclosure Limitation of Establishment Tabular Data," Journal of Official Statistics 14 (4) (1998): 537-551.
[29] J.P. Reiter, "Estimating Probabilities of Identification for Microdata," Journal of the American Statistical Association,,100 (472) (2005): 1103-1112.
[30] Personal correspondence with William E. Winkler, U.S. Census Bureau, September 9, 2005.
[31] U.S. GAO, "Record Linkage and Privacy."

To some, synthetic data is a fancy title for made-up data. In fact, synthetic data sets are not made up at all. They are reliable reproductions of the properties underlying the confidential source file. Synthetic data have essentially the same distribution of characteristics as the source data. Because the data are synthetic, the identity of individuals or businesses is protected. However, the data are not completely safe from re-identification.

The important point for data users is that synthetic data are high-quality data that broaden access at relatively low cost while protecting the identity of individuals and businesses. The Census Bureau can release, for example, more age categories and more detailed geography than would be possible with traditional confidentiality techniques.[32]

Synthetic data allow the same types of statistical analyses that were possible using older but less effective confidentiality techniques.[33] Synthetic data result in clearer analyses of events than do data sets that are distorted by the traditional techniques because no biases are introduced, leading to incorrect conclusions.[34] As Abowd says, "The difference between synthetic data techniques and conventional confidentiality protection methods is that the synthetic data can be designed to minimize bias, relative to the gold standard estimate. The major sources of bias are, then, the same ones that applied to the confidential data--namely, nonsampling or frame biases. These will generally infect both the synthetic and gold standard confidential data."[35]

An enormous advantage is that a source file can release synthetic data from multiple public use files that are tailored to provide different levels of detail. For example, one data user, such as a transportation agency, may need geographic detail, while an economic development agency may need detailed tabulations on industry categories. Data users have always been forced to make a trade-off between the two. With synthetic data, it is possible to produce two public use micro data sets, one with geographic detail and a second with detailed tabulations.

Synthetic data are generally created by sequential regression imputations, one variable in one record at a time. As Zayatz describes, researchers use all the original data and develop a regression model for a given variable. For each record, the researcher blanks the value of that variable and uses the model to impute for it. The process is repeated for each variable.[36]

---

[32] John M. Abowd and Julia I. Lane, "The Economics of Data Confidentiality," October 16, 2003, based on a speech delivered by Lane at the Conference of European Statisticians in Geneva Switzerland (June 12, 2003), and a presentation by Abowd and Lane to the National Science Foundation Workshop on Data Confidentiality, May 11, 2003.

[33] Julia Lane, "Synthetic Data and Confidentiality Protection," Local Employment Dynamics Technical Paper no. TP-2003-10 (Washington: U.S. Census Bureau, Longitudinal Employer-Household Dynamics, 2006, p. 1): available at http://lehd.dsd.census.gov/.

[34] Abowd and Lane, "The Economics of Data Confidentiality."

[35] Personal correspondence with John Abowd, Cornell University, August 31, 2005.

[36] Laura Zayatz, "Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update." RRS 2005/06 (Washington: U.S. Census Bureau, Statistical Research Division, 2006, p. 10): available at www.census.gov/srd/papers/pdf/rrs2005-06.pdf.

Statisticians have developed methods to create partially synthetic data or fully synthetic data. To create partially synthetic data, both demographic and establishment, researchers synthesize a targeted subset of variables for a subset of records that are most likely to cause disclosure.[37] Partially synthetic data replace some or all of the data items on the original source records with synthetic values derived by sampling from an appropriate probability model.

Fully synthetic data, in contrast, synthesize all variables for all records. Fully synthetic data are the result of a complicated method of creating a public use file from synthetic samples of the population, that is, "samples created by taking all the sampling frame variables and generating synthetic individuals that have the same characteristics as the original sample of interest but are not in fact real people."[38] Reiter describes fully synthetic data as fitting models with the original survey data.[39]

Re-identification is possible in both types of synthetic data although the risk of disclosure is greater for partially synthetic data sets than for fully synthetic data sets.

In addition to confidentiality issues, researchers want data sets to have the same statistical properties as the original data ("analytic validity"). That is, the distributions, the means, and the relationships among the variables would ideally be essentially the same as in the actual data file. [40]

As Reiter, Feinberg, and Winkler have all noted, there is a trade-off between the degree of confidentiality and analytic properties.[41] Reiter refers to "refining" the model to achieve extra analytic properties in the masked microdata, and to "coarsening" the model to provide better confidentiality protection.

Differences between the synthetic methods include the degree of their complexity. The partially synthetic data set also includes some actual records and, therefore, added risk of disclosure. Reiter points out that the resulting microdata from both methods can be analyzed using standard statistical techniques and software, albeit with formulas that are dependent on the methods that were used to synthesize the data. In both cases, data users receive data sets in the same

---

[37] Ibid., p. 11.
[38] Ibid.
[39] Jerome P. Reiter, "Releasing Multiply Imputed Synthetic Public Use Microdata: An Illustration and Empirical Study," Journal of the Royal Statistical Society, Series A, 168 (2005): 185.
[40] Abowd and coauthors note that the goal is for the data themselves to determine the nature of the relationships, not the outcome of previous research or the beliefs of the researcher. Once the data set is created, however, it must be tested against previous research to test the validity of the procedure. John Abowd, Gary Benedetto, and Martha H. Stinson, "The Covariance of Earnings and Hours Revisited." Paper presented at SOLE/ EALE conference, San Francisco, June 2005, p. 15.
[41] S.E. Fienberg, "Confidentiality and Disclosure Limitation Methodology: Challenges for National Statistics and Statistical Research" (Washington: Committee on National Statistics of the National Academy of Sciences, 1997); W.E. Winkler, "Masking and Re-identification Methods for Public-Use Microdata: Overview and Research Problems." In J. Domingo-Ferrer and V. Torra, eds., Privacy in Statistical Database (New York: Springer-Verlag, 2004), pp. 231-247, available at www.census.gov/srd/papers/pdf/rrs2004-06.pdf.

format that they would receive if they were using conventional microdata public use files from a sample of actual records.

The method Abowd developed to use behind the Census Bureau's firewall improves synthetic data by "drawing from the posterior predictive distribution conditional on confidential data."[42] This approach helps provide a better match between the synthetic and the original confidential data.

Zayatz at the Census Bureau suggests that, "If we want to begin releasing public use files that link our data with data from other agencies, synthetic data are probably our only choice. Other statistical avoidance techniques are not sufficient to protect the confidentiality of such files."[43] The level of confidentiality afforded to combined files is still a question, as is the analytic validity. Abowd, Bendetto, and Stinson have been testing the analytic and statistical validity of their methodology for partially synthetic earnings data on a SIPP-SSA-IRS file. The eventual goal is for the Census Bureau to release it as a public use file. The synthetic data were compared with the edited version of the original linked file.[44] Work is also underway to produce a partially synthetic data file for public use from an integrated longitudinal business database. According to Winkler, "Abowd's method of iterative deletion and imputation under a suitable model is not fully rigorous. Abowd does give more suitable details that serve as an improvement over previous methods that are similar."[45]

The Census Bureau's Disclosure Review Board (DRB) recently approved its first data product—a set of maps of transportation data developed by Abowd—based on partially synthetic data. The DRB determined that the synthetic data were sufficiently different from the original data, especially in small geographic areas.[46] In this case, the product synthesizes only a small number of variables.

Reiter recently performed simulations on the Current Population Survey that resulted in the release of fully synthetic microdata. He concludes it is necessary to release or describe the imputation models along with the synthetic data created from the models. Given the growing concerns about confidentiality, Reiter recommends using multiple models on the same data set to

---

[42] Personal correspondence with John Abowd, Cornell University July 8, 2005. A method to produce lower-quality synthetic data outside the Census Bureau's firewall is described by Lane, "Synthetic Data and Confidentiality Protection"; and Julia Lane, "Key Issues in Confidentiality Research: Results of an NSF Workshop" (Washington: National Science Foundation, May 12, 2003); and Abowd and Lane, "The Economics of Data Confidentiality.

[43] Zayatz, "Disclosure Avoidance Practice and Research."

[44] The partially synthetic data from the linked IRS detailed earnings records, the Social Security benefit data, and the SIPP records were created "using the structure of the existing SIPP panels with all data elements synthesized using Bayesian bootstrap and sequential regression multivariate imputation methods." John M. Abowd, "Synthetic Data: A New Future for Public Use Micro-Data?" Presentation December 7, 2004 available from http://lehd.dsd.census.gov/led/library/presentations/Synthetic-Data-Census-20041207.pdf

[45] Personal communication with William Winkler, U.S. Census Bureau, September 2005.

[46] Zayatz, "Disclosure Avoidance Practice and Research," p. 11.

better understand their effects on the original data and their value in protecting confidentiality while also providing ease of use for a wide variety of data users.[47]

With synthetic data methods, researchers can measure the effect of disclosure protection on the data set, but the quality of inferences from the synthetic data depends on imputation models."[48] Lane suggests that researchers can further test the data quality of the synthetic data using the source data at the Census Bureau's research data centers.

Winkler's point is worth noting that today's methods for record linkage are very powerful and sophisticated, and that re-identification, while difficult, is possible when using equally sophisticated record linkage techniques.[49] The more variables in the analysis, the more information is available for re-identification and the less likely confidentiality can be ensured. Winkler argues that using more than six variables may compromise confidentiality.[50] If files that use traditional techniques of additive noise have a re-identification rate of 1 percent or more, the Census Bureau uses additional procedures, such as mixtures of additive noise, to reduce that rate.[51]

The Census Bureau expects to add synthetic data techniques to the LED QWI data files to remove the current suppressions for certain of the indicators that are based on small counts. Partially synthetic data offer the LED program a powerful safeguard for avoiding disclosure and at the same time, meet the twin objectives of access to data and protection of the information provided by individuals or business.

---

[47] Reiter, "Releasing Multiply Imputed Synthetic Public Use Microdata, p. 200. See also Reiter, "Inference for Partially Synthetic, Public Use Microdata Sets," Survey Methodology 29 (2) (2003): 181-188.

[48] T.E. Raghunathan, J.P. Reiter, and D.B. Rubin, "Multiple Imputation for Statistical Disclosure Limitation,"Journal of Official Statistics, 19 (1) (2003): 14. Also see J.P. Reiter and T.E. Raghunathan, "Multiple Imputation for Missing Data in Surveys with Complex Designs," Technical Report (Durham, NC: Duke University, Institute of Statistics and Decision Sciences, 2002).

[49] William E. Winkler, "Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata." RRS 2005/09 (Washington: U.S. Census Bureau, Statistical Research Division, October 3, 2005, p. 3): available at www.census.gov/srd/papers/pdf/rrs2005-09.pdf. Other reports by Winkler are listed in Appendix B and some are available at: www.census.gov/srd/www/byname.html.

[50] Ibid., p. 13.

[51] Winkler prefers mixtures of additive noise to techniques such as swapping which degrades data quality more. See Winkler, "Re-identification Methods," and J. Kim, and W. E. Winkler, "Multiplicative Noise for Masking Continuous Data," Proceedings of the Survey Research Methods Section (Washington: American Statistical Association, 2001, CD-ROM); W. Yancey, W.E. Winkler, and R. Creecy, "Disclosure Risk Assessment in Perturbative Microdata Protection." In J. Domingo-Ferrer, ed., Inference Control in Statistical Databases (New York: Springer-Verlag, 2002).

V. IMPLICATIONS FOR NICS

The encouragement by NICS for local and state governments to make more administrative records available for statistical purposes will be of enormous benefit to research and to government and businesses that rely on such research in shaping issues and policies that affect the quality of life and prosperity of communities. Yet, with greater availability comes greater risk that individuals will be identified. In response to this heightened risk, it is likely that federal statistical agencies will further restrict, if not eventually abandon, the release of microdata public use files that use only traditional tools to protect confidentiality.

For NICS, it is clearly beneficial to educate data owners about tools and the options for disclosure avoidance, their benefits, and risks. For example, access to the original data through remote submission of data requests (offline and online) is a reasonable option for NICS to support and promote, given that it is adequate for many less complex research projects. The procedures and software exist already.

A. Traditional Confidentiality Tools

The choice of confidentiality techniques affects both data availability and data quality in different ways.[52] Most traditional tools are impractical or inadequate for local and state public use microdata files.

Research is underway to determine if Abowd's method of multiplicative noise infusion in conjunction with other traditional methods provides a higher level of confidentiality assurance than traditional methods alone.[53] Abowd's is a sophisticated method that includes confidential factors and methodology that are not immediately available to local areas. If these methods are implemented at the start of a data project, however, they can be used consistently throughout its life, potentially justifying the investment.

B. Synthetic Microdata

Synthetic microdata data files meet the objectives of avoiding re-identification of individuals better than traditional techniques, providing high-quality data, and being practical for state and larger substate governments to apply to their data sets. There are, however, significant development issues before this methodology is NICS-ready.

---

[52] A. Dobra, S. E Fienberg, and M. Trottini, "Assessing the Risk of Disclosure of Confidential Categorical Data." In J. M. Bernardo and others, eds., Bayesian Statistics, vol. 7 (New York: Oxford University Press, 2003).

[53] Sam Hawala, Martha Stinson, and John Abowd, "Disclosure Risk Assessment Through Record Linkage," Joint UNECE/Eurostat work session on statistical data confidentiality (Geneva, Switzerland, November 9-11, 2005). The authors conclude that there was both good news and bad news from the preliminary results of their research on disclosure risks in a proposed public use file with synthetic data. There were cells with a disproportionate share of true matches. They are continuing work to consider strategies that balance the need for high data quality while also reducing the risk of disclosure.

Fully synthetic data are the best at avoiding disclosure but are probably too methodologically complex to expect local and state governments to use. As Abowd notes, "The real problem with fully synthetic data for a locality is that they require a good frame (like the Census of Population or the Economic Census) from which the synthetic samples are drawn."[54] That leaves partially synthetic data as the method most appropriate for further development by NICS.

C. Options for NICS

In the short term, helping data producers understand and use traditional tools, along with tools that can be used on the internet, such as automatically aggregating geographic areas when there is a risk of disclosure, is the most reasonable option for NICS. Many of the uses at the local level are for small geographic areas, which raise issues of how to protect data confidentiality while maintaining data quality, and these issues are often different from those federal agencies face.

Over a longer term, however, traditional tools are inadequate, especially given the proliferation on the Internet of data files with personal information. The pressing need is for tools to create synthetic data sets.

The methodology for creating synthetic data is currently being developed under National Science Foundation (NSF) grants. What is needed is practical, affordable application software that (1) converts actual administrative records to partially synthetic data; and (2) checks the statistical validity of the partially synthetic data against the gold standard, the original data file.

Such software, however, is probably a long way off, if it can ever be done. It may be that the NSF or a foundation would fund further development of the methodology for use with smaller administrative files from government programs.

This recommendation does not apply to local or state sample surveys linked to administrative records for conversion to public use microdata files. With sample surveys as part of a linked data set, weighting issues arise in addition to confidentiality issues. Although formulas and implementation methodology exist to address that problem, further work is needed to address confidentiality concerns.

[54] Personal correspondence with John Abowd, Cornell University, August 3, 2005.

VI. IMPLICATIONS FOR THE FEDERAL STATISTICAL SYSTEM

The federal statistical system is continuing its efforts to provide useful public use microdata files for research that also have an acceptably low risk of re-identification of individuals. Two technological advances of recent years have added to the complexity of this issue:

1. A great deal of information is available about individuals on the Internet, from public records such as California's birth and death records, to Kentucky's marriage records, to genealogical sites, and the data mining provided by businesses for a fee.

2. Sophisticated software has been developed that allows records to be linked in ways that can identify a small percentage (perhaps 1 percent) of those in traditional public use microdata files with as few as seven variables.

The reality is that the risks of disclosing individual information have increased significantly in just the last few years. To combat this, federal statistical agencies have taken additional measures to protect confidentiality. For example, the Census Bureau has added noise to PUMS files. Adding noise, however, can negatively affect the use of the PUMS for sophisticated analyses. The Census Bureau has also announced plans to release the American Community Survey (ACS) PUMS as a 1 percent sample rather than a 2.5 percent sample, which increases uncertainty in sample estimates. In addition to "aging" the data, they have added more noise than was the case in the past. Users countered that the ACS plan was unacceptable because a 1 percent sample increases sampling error. Further, it would take five years to accumulate a 5 percent sample. As Julie Hoang of the California State Data Center said of rapidly changing areas, "a five-year aggregation of a 1 percent sample will not validly represent the most recent population. In fact, policy decisions based on these data could be very misguided." Hoang further commented that, "the effects of local government policies need to be evaluated over a short time frame. A five-year time frame will not permit us to identify whether a particular policy intervention is associated with a certain outcome at the local level."[55]

Data users are perplexed by what seems to be unreasonable actions because, understandably, they are not well versed on the recent advances in record linking technology. Federal statistical agencies have been focusing on new threats to confidentiality, but they must increase both discussion of the topic and the audience that hears their concerns about disclosures. There is an extensive literature to support the legitimacy of the concerns of the federal statistical agencies (see Appendix B).

We believe that statistical agencies should invest even more in finding alternatives to public use microdata as the primary way that a broad group of researchers can produce customized tabulations. For example:

[55] E-mail correspondence between Julie Hoang, California State Data Center, and Lisa Blumerman, Deputy Chief of the American Community Survey, May 1, 2006.

- Abowd's work to produce synthetic estimates deserves further support, both from statistical agencies in making the method feasible for data sets such as the American Community Survey, and from data users, who should to learn how to use these data in their research.
- The NASS system of automatically aggregating geography where the risk of disclosure is too great could be further developed, perhaps in conjunction with remote submission of tabulation programs.
- Further development of technology to dynamically assess disclosure risk could help statistical agencies feel more secure about the problems associated with the release of special tabulations over time and across geographic areas. Such information may also educate data users about the very real risks of disclosure with which statistical agencies must contend.

For these alternatives to be developed, however, statistical agencies will require additional funding.

APPENDIX A

MICRODATA CONFIDENTIALITY REFERENCES

Compiled by William E. Winkler, U.S. Census Bureau (william.e.winkler@census.gov)
June 21, 2005

Abowd, J. M., and S. D. Woodcock, S. D. 2002. "Disclosure Limitation in Longitudinal Linked Data." In P. Doyle and others, eds., *Confidentiality, Disclosure, and Data Access.* Amsterdam: North Holland.

———. 2004. "Multiply-Imputing Confidential Characteristics and File Links in Longitudinal Linked Data." In J. Domingo-Ferrer and V. Torra, eds., *Privacy in Statistical Databases 2004.* New York: Springer.

Adams, N. R., and J. C. Wortmann. 1989. "Security-control Methods for Statistical Databases: A Comparative Study." *ACM Computing Surveys* 21: 515-556.

Aggarwal, G., and others. 2005. "Anonymizing Tables," *Proceedings of the 10th International Conference on Database Theory*. Edinburgh, Scotland.

Agrawal, D., and C. C. Aggarwal. 2001. "On the Design and Quantification of Privacy Preserving Data Mining Algorithms." Association of Computing Machinery, Proceedings of PODS (2001): 247-255.

Agrawal, R., and R. Srikant. 2000. "Privacy Preserving Data Mining," In Weidong Chen, Jeffrey F. Naughton, and Philip A. Bernstein, eds., Proceedings of the ACM SIGMOD 2000. New York: Association for Computing Machinery.

Agrawal, R., Srikant, R., and D. Thomas, D. 2005. "Privacy Preserving OLAP." Paper presented at the ACM SIGMOD Conference, Baltimore, MD, June 14-16.

Agrawal, R., and others. 2002. "Hippocratic Databases," *Very Large Databases* (2002).

Bacher, J., S. Bender, and R. Brand, R. 2001. "Re-identifying Register Data by Survey Data: An Empirical Study." Presented at the UNECE Workshop on Statistical Data Editing, Skopje, Macedonia, May.

Bacher, J., R. Brand, and S. Bender. 2002. "Re-identifying Register Data by Survey Data Using Cluster Analysis: An Empirical Study." *International Journal of Uncertainty, Fuzziness, Knowledge-Based Systems* 10 (5): 589-608.

Bayardo, R. J., and R. Agrawal. 2005. "Data Privacy Through Optimal K-Anonymization." IEEE 2005 International Conference on Data Engineering. Washington: IEEE Computer Society.

Bethlehem, J. A., W. J. Keller, and J. Pannekoek. 1990. "Disclosure Control of Microdata." *Journal of the American Statistical Association* 85: 38-45.

Blien, U., U. Wirth, and M. Muller. 1992. "Disclosure Risk for Microdata Stemming from Official Statistics." *Statistica Neerlandica* 46: 69-82.

Blum, A., and others. 2005. "Practical Privacy: The SuLQ Framework." ACM SIGMOD Conference, Baltimore, MD, June 14-16, available at http://research.microsoft.com/research/sv/DatabasePrivacy/bdmn.pdf.

Brand, R. 2002. "Microdata Protection Through Noise Addition." In J. Domingo-Ferrer, ed., Inference Control in Statistical Databases. New York: Springer.

Castro, J. 2004. “Computational Experience with Minimum-Distance Controlled Perturbation Methods.” In J. Domingo-Ferrer and V. Torra, ed., *Privacy in Statistical Databases 2004.* New York: Springer.

Chawla, S., and others. 2004. “Toward Privacy in Public Databases.” Microsoft Research Technical Report. Theory of Cryptography Conference.

Chawla, S., and others. 2005. “On the Utility of Privacy-Preserving Histograms,” available at http://research.microsoft.com/research/sv/DatabasePrivacy/cdmt.pdf .

Dalenius, T., and S.P. Reiss. 1982. “Data-swapping: A Technique for Disclosure Control.” Journal of Statistical Planning and Inference 6: 73-85.

Dandekar, R. A. 2004. “Maximum Utility Minimum Information Loss Table Server Design of Statistical Disclosure Control of Tabular Data.” In J. Domingo-Ferrer and V. Torra, eds., *Privacy in Statistical Databases.* New York: Springer.

Dandekar, R. A., J. Domingo-Ferrer, and F. Sebe. 2002. “LHS-Based Hybrid Microdata vs Rank Swapping and Microaggregation for Numeric Microdata Protection.” In J. Domingo-Ferrer and V. Torra, eds., *Inference Control in Statistical Databases.* New York: Springer.

Dandekar, R., M. Cohen, and N. Kirkendal. 2002. “Sensitive Microdata Protection Using Latin Hypercube Sampling Technique.” In J. Domingo-Ferrer and V. Torra, eds., *Inference Control in Statistical Databases.* New York: Springer.

Defays, D., and M. N. Anwar. 1998. “Masking Microdata Using Micro-aggregation.*” Journal of Official Statistics* 14: 449-461.

Defays, D., and P. Nanopolis, P. 1993. “Panels of Enterprises and Confidentiality: The Small Aggregates Method.” In *Proceedings of the 1992 Symposium on Design and Analysis of Longitudinal Surveys*, 195-204.

De Waal, A. G., and L. Willenborg. 1995. "Global Recodings and Local Suppressions in Microdata Sets." *Proceedings of Statistics Canada Symposium 95*, 121-132.

———. 1996. "A View of Statistical Disclosure Control for Microdata." *Survey Methodology* 22: 95-103.

Dinur, I., and K. Nissim. 2003. “Revealing Information while Preserving Privacy.” Paper presented at the ACM SIGMOD/PODS Conference, San Diego, CA, June 9-12.

Domingo-Ferrer, J. (2001). “On the Complexity of Microaggregation.” Paper presented at the UNECE Workshop on Statistical Data Editing, Skopje, Macedonia, May.

——— ed. 2002. *Inference Control in Statistical Databases*, New York: Springer

Domingo-Ferrer, J., and J. M. Mateo-Sanz. 2001. “An Empirical Comparison of SDC Methods for Continuous Microdata in Terms of Information Loss and Re-Identification Risk.” Paper presented at the UNECE Workshop On Statistical Data Editing, Skopje, Macedonia, May.

———. 2002.“Practical Data-Oriented Microaggregation for Statistical Disclosure Control.” *IEEE Transactions on Knowledge and Data Engineering* 14 (1): 189-201.

Domingo-Ferrer and others. 2002. “On the Security of Microaggregation with Individual Ranking: Analytic Attacks.*” International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems* 10 (5): 477-492.

Domingo-Ferrer, J., F. Sebé, and J. Castellà-Roca. 2004. “On the Security of Noise Addition for Privacy in Statistical Databases. In J. Domingo-Ferrer and V. Torra, eds., *Privacy in Statistical Databases.* New York: Springer.

Domingo-Ferrer, J., and V. Torra. 2001. “A Quantitative Comparison of Disclosure Control Methods for Microdata.” In P. Doyle and others, eds., *Confidentiality, Disclosure Control and Data Access: Theory and Practical Applications*. Amsterdam: North Holland.

———. 2003. “Statistical Data Protection in Statistical Microdata Protection via Advanced Record Linkage.” *Statistics and Computing* 13 (4): 343-354.

Du., W., Y. Han, and S. Chen. 2004. “Privacy Preserving Multivariate Statistical Analysis: Linear Regression and Classification.” SIAM International Conference on Data Mining. Lake Buena Vista, FL, April 22-24.

Du, W., and Z. Zhan. 2003. “Using Randomized Response Techniques for Privacy Preserving Data Mining.” In Lise Getoor and others, eds., Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: Springer.

DuMouchel, W., and others. 1999. “Squashing Flat Files Flatter.” Proceedings of the ACM *Knowledge Discovery and Data Mining Conference.* New York: Association of Computing Machinery.

Duncan, G. T., S. Keller-McNulty, and S. Stokes. 2001. “Disclosure Risk vs. Data Utility: The R-U Confidentiality Map.” Los Alamos National Laboratory Technical Report LA-UR-01-6428. Los Alamos, NM: Los Alamos National Laboratory.

Dwork, C., and K. Nissim. 2004. “Privacy-Preserving Data Mining on Vertically Partitioned Databases.” Microsoft Research Technical Report.

Elamir, E. A. H. 2004. “Analysis of Re-identification Risk based on Log-Linear Model.” In J. Domingo-Ferrer and V. Torra, eds., *Privacy in Statistical Databases.* New York: Springer.

Elliott, M. A., A.M. Manning, and R. W. Ford. 2002. “A Computational Algorithm for Handling the Special Uniques Problem.” *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems* 10 (5): 493-510.

Elliot, M.J., C.A. Skinner, and A. Dale. 1998. “Special Uniques, Random Uniques, and Sticky Populations: Some Counterintuitive Effects of Geographic Detail in Disclosure Risk.” Research in Official Statistics 1: 53-68.

Evfimievski, A. 2004. “Privacy Preserving Information Sharing,” Ph.D. Dissertation, Cornell University, available at www.cs.cornell.edu/aevf/ .

Evfimievski, A., J. Gehrke, and R. Srikant. 2003. “Limiting Privacy Breaches in Privacy Preserving Data Mining.” In Alon Y. Halevy, Zachary G. Ives, and AnHai Doan, eds., Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data. New York: Association of Computing Machinery.

Evfimievski, A., Srikant, R., Agrawal, R., and Gehrke, J. 2002. “Privacy Preserving Mining of Association Rules.” New York: Association of Computing Machinery, Special Interest Group on Knowledge Discovery and Datamining.

Fellegi, I. P. 1972. "On the Question of Statistical Confidentiality." *Journal of the American Statistical Association* 67: 7-18.

———. 1999. “Record Linkage and Public Policy - A Dynamic Evolution.” *Proceedings of the Record Linkage Workshop 1997.* Washington: National Academy Press.

Fellegi, I. P., and A. B. Sunter. 1969. "A Theory for Record Linkage." *Journal of the American Statistical Association* 64: 1183-1210.

Fienberg, S. E. 1997. “Confidentiality and Disclosure Limitation Methodology: Challenges for National Statistics and Statistical Research.” Paper commissioned by Committee on National Statistics of the National Academy of Sciences.

Fienberg, S. E., E. U. Makov, and A. P. Sanil. 1997. “A Bayesian Approach to Data Disclosure: Optimal Intruder Behavior for Continuous Data.” *Journal of Official Statistics* 14: 75-89.

Fienberg, S. E., E. U. Makov, and R. J. Steel. 1998. “Disclosure Limitation Using Perturbation and Related Methods for Categorical Data.” *Journal of Official Statistics* 14: 485-502.

Frakes, W., and Baeza-Yates, R. 1992. *Information Retrieval - Data Structures and Algorithms.* Upper Saddle River, NJ: Prentice-Hall.

Franconi, L. and others. 2001. “Experiences in Model-Based Disclosure Protection.” Paper presented at the UNECE Workshop on Statistical Data Confidentiality, Skopje, Macedonia, May.

Fuller, W. A. 1993. “Masking Procedures for Microdata Disclosure Limitation.” Journal of Official Statistics 9: 383-406, available at www.jos.nu/Articles/abstract.asp?article=92383.

Fung, B. C. M., K. Wang, K., and P.S. Yu. 2005. “Top-Down Specialization for Information and Privacy Preservation.” IEEE International Conference on Data Engineering, September 2004,. Washington: IEEE Computer Society.

Gopal, R., P. Goes, and R. Garfinkel. 1998. “Confidentiality Via Camouflage: The CVC Approach to Database Query Management.” In *Statistical Data Protection ’98.* Brussels: Eurostat.

Gilburd, B., A. Schuster, and R. Wolff. 2004. “k-TTP: A New Privacy Model for Large-Scale Distributed Environments.” *ACM Knowledge Discovery and Data Mining Conference* 2004, 599-604. New York: ACM Press.

Gill, L. 1999. “OX-LINK: The Oxford Medical Record Linkage System.” In *Record Linkage Techniques 1997.* Washington: National Academy Press.

Gomatam, S. V., and A. Karr. 2003. “On Data Swapping of Categorical Data.” *Proceedings of the Survey Research Methods Section.* Washington: American Statistical Association, CD-ROM.

Gouweleeuw, J.M. and others. 1998. “Post Randomisation For Statistical Disclosure Control: Theory and Implementation.” *Journal of Official Statistics* 14: 463-478.

Grim, J., P. Bocek, and P. Pudil. 2001. “Safe Dissemination of census Results by Means of Interactive Probabilistic Models.” *Proceedings of 2001 NTTS and ETK.* Luxembourg: Eurostat.

Huang, Z., W. Du, and B. Chen. 2005. “Deriving Private Information from Randomized Data,” In Fatma Ozcan, ed., Proceedings of the ACM SIGMOD International Conference on Management of Data. New York: ACM.

Hwang, J. T. 1986. “Multiplicative Error-in-Variables Models with Applications to Recent Data Released by the U.S. Department of Energy.” Journal of the American Statistical Association 81 (395): 680-688.

Iyengar, V. 2002. “Transforming Data to Satisfy Privacy Constraints.” Association of Computing Machinery, Knowledge Discovery and Data Mining Conference 2002. New York: ACM.

Kennickell, A. B. 1999. “Multiple Imputation and Disclosure Control: The Case of the 1995 Survey of Consumer Finances.” In Record Linkage Techniques 1997. Washington: National Academy Press, available at http://www.fcsm.gov under methodology reports.

Kantarcioglu, M., and C. Clifton. 2004. “Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data.” *IEEE Transactions on Knowledge and Data Engineering* 16 (9): 1026-1037.

———. 2004. “When Do Data Mining Results Violate Privacy?” *Association of Computing Machinery, Knowledge Discovery and Data Mining Conference* 2004. New York: ACM.

Kargupta, H. and others. 2003. “Random Data Perturbation Techniques and Privacy Preserving Data Mining.” Expanded version of best paper awarded paper from the IEEE International Conference on Data Mining, November, 2003, Orlando, FL (also to appear in *Knowledge and Information Systems Journal (*www.cs.umbc.edu/~hillol/PUBS/kargupta_privacy03a.pdf)).

Kim, J. J. 1986. "A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation." Proceedings of the Survey Research Methods Section. Washington: American Statistical Association, available at www.amstat.org/sections/SRMS/Proceedings/papers/1986_069.pdf.

———. 1990. "Subdomain Estimation for the Masked Data." Proceedings of the Section on Survey Research Methods. Washington: American Statistical Association, available at: www.amstat.org/sections/SRMS/Proceedings/papers/1990_075.pdf.

Kim, J. J., and W. E. Winkler. 1995. “Masking Microdata Files,” Proceedings of the Survey Research Methods Section. Washington: American Statistical Association, available at: www.amstat.org/sections/SRMS/Proceedings/papers/1995_017.pdf.

———. 2001. “Multiplicative Noise for Masking Continuous Data,” Proceedings of the Survey Research Methods Section. Washington: American Statistical Association, CD-ROM.

Lambert, D. 1993. “Measures of Disclosure Risk and Harm,” Journal of Official Statistics, 9: 313-331, available at www.jos.nu/Articles/abstract.asp?article=92313.

Lawrence, C., J. L. Zhou, and A. L. Tits. 1997. “User’s Guide for CFSZP Version 2.5: A C Code for Solving (Large Scale) Constrained Nonlinear Inequality Constraints.” Unpublished manuscript, Electrical Engineering Dept. and Institute for Systems Research, University of Maryland.

Lakshmanan, L., R. Ng, and G. Ramesh 2005. “To Do or Not To Do – The Dilemma of Disclosing Anonymized Data.” Paper presented at the ACM SIGMOD Conference, Baltimore, MD, June 14-16.

LeFevre, K., D. DeWitt, and R. Ramakrishnan. 2005. “Incognito: Efficient Full-Domain K-Anonymity.” Paper presented at the ACM SIGMOD Conference, Baltimore, MD, June 14-16.

Liew, C. K., U. J. Choi, and C. J. Liew. 1991. “A Data Distortion by Probability Distribution.” *ACM Transactions on Database Systems* 10: 395-411.

Little, R. J. A. 1993. “Statistical Analysis of Masked Data.” Journal of Official Statistics 9: 407-426, available at www.jos.nu/Articles/abstract.asp?article=92407.

Little, R. J. A., and F. Liu. 2002. “Selective Multiple Imputation of Keys for Statistical Disclosure Control in Microdata,” *Proceedings of the Survey Research Methods Section.* Washington: American Statistical Association, CD-ROM.

———. 2003. “Comparison of SMIKe with Data-Swapping and PRAM for Statistical Disclosure Control of Simulated Microdata,” *Proceedings of the Survey Research Methods Section.* Washington: American Statistical Association, CD-ROM.

Lindell, Y., and B. Pinkas. 2002. “Privacy Preserving Data Mining,” *Proceedings of Crypto 2000.* New York: Springer LNCS 1880.

Malin, B. L. Sweeney, and E. Newton. 2003. “Trail Re-identification: Learning Who You are from Where You have Been,” Workshop on Privacy in Data, Carnegie-Mellon University, March 2003.

Mera, R. 1998. “Matrix Masking Methods That Preserve Moments.” *Proceedings of the Survey Research Methods Section.* Washington: American Statistical Association.

Moore, R. 1995. “Controlled Data Swapping Techniques For Masking Public Use Data Sets.” Report rr96/04. Washington: U.S. Bureau of the Census, Statistical Research Division, available at http://www.census.gov/srd/www/byyear.html.

Moore, A. W., and M. S. Lee. 1998. “Cached Sufficient Statistics for Efficient Machine Learning with Large Datasets.” *Journal of Artificial Intelligence Research*, 8: 67-91.

Müller, W., U. Blien, and H. Wirth, H. 1995. “Identification Risks of Micro Data: Evidence from Experimental Studies.” *Sociological Methods and Research* 24: 131-157.

Muralidhar, K., D. Batrah, and P.J. Kirs. 1995. “Accessibility, Security, and Accuracy in Statistical Databases: The Case for the Multiplicative Fixed Data Perturbation Approach.” *Management Science* 41(9): 1549-1584.

Muralidhar, K., R. Parsa, and R. Sarathy. 1999. “A General Additive Data Perturbation Method for Database Security.” *Management Science* 45 (10): 1399-1415.

Muralidhar, K., and R. Sarathy. 1999. "Security of Random Data Perturbation Methods." *ACM Transactions on Database Systems* 24 (4): 487-493.

———. 2003. “A Theoretical Basis for Perturbation Methods.” *Statistics and Computing* 13 (4): 329-335.

Muralidhar, K., R. Sarathy, and R. Parsa. 2001. "An Improved Security Requirement for Data Perturbation with Implications for E-Commerce." *Decision Sciences* 32 (4): 683-698.

Owen, A. 2003. “Data Squashing by Empirical Likelihood,” *Data Mining and Knowledge Discovery* 7 (1): 101-113.

Paas, G. 1988. "Disclosure Risk and Disclosure Avoidance for Microdata." *Journal of Business and Economic Statistics* 6: 487-500.

Palley, M. A., and J. S. Simonoff. 1987. “The Use of Regression Methodology for the Compromise of Confidential Information in Statistical Databases.” *ACM Transactions on Database Systems* 12 (4): 593-608.

Polletini, S. 2003. “Maximum Entropy Simulation for Microdata Protection.” *Statistics and Computing* 13 (4): 307-320.

Polletini, S., L. Franconi, and J. Stander. 2002. “Model Based Disclosure Protection.” In J. Domingo-Ferrer, ed., *Inference Control in Statistical Databases.* New York: Springer.

Polletini, S., and J. Stander. 2004. “A Bayesian Hierarchical Model Approach to Risk Estimation in Statistical Disclosure Limitation.” In J. Domingo-Ferrer and V. Torra, eds., *Privacy in Statistical Databases.* New York: Springer.

Raghunathan, T. E. 2003. “Evaluation of Inferences from Multiple Synthetic Data Sets Created Using Semiparametric Approach, Panel on Confidential Data Access for Research Purposes.” Washington: Committee on National Statistics.

Raghunathan, T. E. and others. 1998. “A Multivariate Technique for Multiply Imputing Missing Values Using a Series of Regression Models.” Ann Arbor, MI: Survey Research Center, University of Michigan.

Raghunathan, T.E., J. P. Reiter, and D. R. Rubin. 2003. “Multiple Imputation for Statistical Disclosure Limitation.” *Journal of Official Statistics* 19: 1-16.

Raghunathan, T.E., and D. R. Rubin. 2000. “Multiple Imputation for Disclosure Limitation.” Technical report. Ann Arbor: University of Michigan, Department of Biostatistics.

Reiss, J.P. 1984. “Practical Data Swapping: The First Steps.” *ACM Transactions on Database Systems* 9: 20-37.

Reiter, J.P. 2002. “Satisfying Disclosure Restrictions with Synthetic Data Sets.” *Journal of Official Statistics* 18: 531-543.

———. 2003. “Inference for Partially Synthetic, Public Use Data Sets.” *Survey Methodology* :181-189.

———. 2003. “Estimating Probabilities of Identification for Microdata.” Washington: Panel on Confidential Data Access for Research Purposes, Committee on National Statistics.

———. 2005. “Releasing Multiply Imputed, Synthetic Public Use Microdata: An Illustration and Empirical Study.” *Journal of the Royal Statistical Society, Series A, page forthcoming.*

Roque, G. M. 2000. “Masking Microdata Files with Mixtures of Multivariate Normal Distributions.” Ph.D.Dissertation, Department of Statistics, University of California at Riverside.

Rubin, D. B. 1993. “Satisfying Confidentiality Constraints through the Use of Synthetic Multiply-imputed Microdata.” *Journal of Official Statistics* 91: 461-468.

Samarati, P. 2001. “Protecting Respondents’ Identity in Microdata Release.” *IEEE Transactions on Knowledge and Data Engineering* 13 (6): 1010-1027.

Samarati, P., and L. Sweeney. 1998. “Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement Through Generalization and Cell Suppression.” Technical Report. New York: SRI International.

Sarathy, R., and K. Muralidhar. 2002. "The Security of Confidential Numerical Data in Databases." *Information Systems Research* 48 (12): 1613-1627.

Sarathy, R., Muralidhar, K., and Parsa, R. (2002), Perturbing Non-Normal Attributes: The Copula Approach.” *Management Science* 48 (12), 1613-1627.

Scheuren, F., and W. E. Winkler. 1993. “Recursive Merging and Analysis of Administration Lists.” *Proceedings of the Survey on Research Methods Section.* Washington: American Statistical Association, available at www.amstat.org in the Section on Government Statistics.

———. 1997. “Regression Analysis of Data Files that are Computer Matched – Part II,” *Survey Methodology* 23 (2) 157-165.

Schlörer, J. 1981. “Security of Statistical Databases: Multidimensional Transformation.” *ACM Transactions on Database Systems* 6: 91-112.

Skinner, C. J., and M. A. Elliot. 2001. “A Measure of Disclosure Risk for Microdata.” *Journal of the Royal Statistical Society* 64 (4): 855-867.

Skinner, C. J., and D. J. Holmes. 1998. “Estimating the Re-identification Risk per Record in Microdata.” *Journal of Official Statistics*14: 361-372.

Stander, J., and L. Franconi, L. 2001. “A Model-Based Disclosure Limitation Method for Business Microdata.” Paper presented at the UNECE Workshop on Statistical Data Editing, Skopje, Macedonia, May.

Sullivan, G., and W. A. Fuller. 1989. "The Use of Measurement Error to Avoid Disclosure." *Proceedings of the Survey on Research Methods Section.* Washington: American Statistical Association.

———. 1990. "Construction of Masking Error for Categorical Variables." *Proceedings of the Survey Research Methods Section*. Washington: American Statistical Association.

Sweeney, L. 1999. “Computational Disclosure Control for Medical Microdata: The Datafly System.” In *Record Linkage Techniques 1997.* Washington: National Academy Press.

———. 2002. “Achieving k-Anonymity Privacy Protection Using Generalization and Suppression.” *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems* 10 (5): 571-588.

———. 2004. “Optimal Anonymity using K-similar, a New Clustering Algorithm,” Unpublished manuscript.

Takemura, A. 2002. “Local Recoding and Record Swapping by Maximum Weight Matching for Disclosure Control of Microdata Sets.” *Journal of Official Statistics* 18 (2): 275-289.

Tendick, P., and N. Matloff. 1994. “A Modified Random Perturbation Method for Database Security.” *ACM Transactions on Database Systems* 19: 47-63.

Thibaudeau, Y. 2004. “An Algorithm for Computing Full Rank Minimal Sufficient Statistics with Application to Confidentiality Protection.” *UNECE Statistical Journal* to appear.

Thibaudeau, Y., and W. E. Winkler. 2002. “Bayesian Networks Representations, Generalized Imputation, and Synthetic Microdata Satisfying Analytic Restraints.” Report RR 2002/09. Washington: U.S. Census Bureau, Statistical Research Division, available at www.census.gov/srd/www/byyear.html.

———. 2004. “Full Rank Minimal Statistics for Disclosure Limitation and Variance Estimation: A Practical Way to Release Count Information.” Proceedings of the Survey Research Methods Section. Washington: American Statistical Association, CD-ROM.

Torra, V. 2004. “OWA Operators in Data Modeling and Re-identification.” *IEEE Transactions on Fuzzy Systems* 12 (5): 652-660.

Torra, V., and S. Miyamoto, S. 2004. “Evaluating Fuzzy Clustering Algorithms for Microdata Protection.” In J. Domingo-Ferrer and V. Torra, eds., *Privacy in Statistical Databases.* New York: Springer.

Trottini, M., and S. E. Fienberg. 2002. “Modeling User Uncertainty for Disclosure Risk and Data Utility.” *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems* 10: 511-528.

Van Den Hout, A., and P. G. M. Van Der Heijden. 2002. “Randomized Response, Statistical Disclosure Control, and Misclassification: A Review.” *International Statistical Review* 70 (2): 269-288.

Van Gewerden, L., A. Wessels, and A. Hundepol. 1997. “Mu-Argus Users Manual, Version 2.” Document TM-1/D. Statistics Netherlands.

Willenborg, L., and T. De Waal. 1996. *Statistical Disclosure Control in Practice*, vol. 111. Lecture Notes in Statistics. New York: Springer.

———. 2000. *Elements of Statistical Disclosure Control*, vol. 155, Lecture Notes in Statistics. New York: Springer.

Winglee, M., and others. 2002. “Assessing Disclosure Protection for the SOI Public Use File,” *Proceedings of the Section on Survey Research Methods.* Washington: American Statistical Association, CD-ROM.

Winkler, W. E. 1994. "Advanced Methods for Record Linkage, Proceedings of the Section on Survey Research Methods. Washington: American Statistical Association, available at: www.census.gov/srd/papers/pdf/rr94-5.pdf.

———. 1995. "Matching and Record Linkage." In B. G. Cox and others, eds., Business Survey Methods. New York: Wiley.

———. 1997. “Views on the Production and Use of Confidential Microdata.” Report RR 97/01. Washington: U.S. Census Bureau, Statistical Research Division, available at www.census.gov/srd/www/byyear.html.

———. 1998. “Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata,” Research in Official Statistics 1:87-104.

———. 2002. “Using Simulated Annealing for k-anonymity.” Report RR 2002/07. Washington: U.S. Census Bureau, Statistical Research Division, available at www.census.gov/srd/www/byyear.html.

———. (2002. “Single Ranking Micro-aggregation and Re-identification.” Report RR 2002/08 (Washington: U.S. Census Bureau, Statistical Research Division): available at www.census.gov/srd/www/byyear.html.

———. 2004. Masking and Re-identification Methods for Public Use Microdata: Overview and Research Problems. In J. Domingo-Ferrer and V. Torra, eds., Privacy in Statistical Database. New York: Springer, available also at www.census.gov/srd/papers/pdf/rrs2004-06.pdf.

———. 2004. Re-identification Methods for Masked Microdata. In J. Domingo-Ferrer and V. Torra, eds., Privacy in Statistical Database. New York: Springer, available also at http://www.census.gov/srd/papers/pdf/rrs2004-03.pdf .

———. 2005. “Modeling Data and Data Quality.” Technical report (, to appear.

———. 2005. “Modeling and Quality of Masked Microdata,” Proceedings of the Survey Research Method Section. Washington: American Statistical Association.

Yancey, W.E., W. E. Winkler, and R. H. Creecy. 2002. “Disclosure Risk Assessment in Perturbative Microdata Protection.” In J. Domingo-Ferrer, ed., Inference Control in Statistical Databases. New York: Springer, available also at www.census.gov/srd/papers/pdf/rrs2002-01.pdf.

Yang, Z., S. Zhong, S., and R. Wright. 2005. “Anonymity Preserving Data Collection,” *Proceedings of the 11th ACM KDD Conference*, August 21-24, Chicago, Ill.

Zhang, N., S. Wang, and W. Zhao. 2005. “A New Scheme on Privacy-Preserving Data Classification.” *Proceedings of the 11th ACM KDD Conference*, August 21-24, Chicago, Ill.Zhong, S., Z. Yang, and R. Wright. 2005. “Privacy-Enhancing k-Anonymization of Customer Data.” Paper presented at the Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems 2005, Baltimore, Maryland, *June 13 - 15, 2005.* New York: ACM Press.

Zhu, Y., and L. Liu. 2004. “Optimal Randomization for Privacy Preserving Data Mining,” *ACM Knowledge Discovery and Data Mining Conference* 2004. New York: ACM Press.

APPENDIX B

RECORD LINKAGE REFERENCES

Compiled by William E. Winkler, U.S. Census Bureau (william.e.winkler@census.gov)
June 21, 2005

Abowd, J., and J. Vilhuber. 2004. "The Sensitivity of Economic Statistics to Coding Errors in Personal Identifiers (with discussion)," *Journal of Business and Economic Statistics*, 23 (2): 158-160.

Agichstein, E., and V. Ganti, V. 2004. "Mining Reference Tables for Automatic Text Segmentation," *ACM Knowledge Discovery and Data Mining Conference 2004*, 20-29. New York: ACM Press.

Alvey, W., B. and Jamerson., eds. 1997. Record Linkage Techniques – 1997: Proceedings of An International Record Linkage Workshop and Exposition, March 20-21, 1997. Washington: American Statistical Association, available at http://www.fcsm.gov/working-papers/RLT_1997.html.

Ananthakrishna, R., S. Chaudhuri, and V. Ganti, V. 2002. "Eliminating Fuzzy Duplicates in Data Warehouse." *Very Large Data Bases* (2002): 586-597.

Armstrong, J. A. 2000. "Weight Estimation for Large Scale Record Linkage Applications." *Proceedings of the Survey Research Methods Section.* Washington: American Statistical Association.,

Armstrong, J. A., C. Block, and M. Saleh. 1999. "Record Linkage for Electoral Administration.", *Proceedings of the Survey Methods Section.* Statistical Society of Canada.

Armstrong, J. A., and J. E. Mayda. 1993. "Model-based Estimation of Record Linkage Error Rates." *Survey Methodology* 19: 137-147.

Battacharya, I., and L. Getoor. 2004. "Iterative Record Linkage for Cleaning and Integration." Paper presented at the ACM SIGMOD Workshop on Data Mining and Knowledge Discovery 2004, Paris, France.

Baxter, R., P. Christen, and T. Churches. 2003. "A Comparison of Fast Blocking Methods for Record Linkage," In *Proceedings of the ACM Workshop on Data Cleaning, Record Linkage and Object Identification.* Washington: ACM.

Belin, T. R. (1993) "Evaluation of Sources of Variation in Record Linkage through a Factorial Experiment." *Survey Methodology* 19, 13-29.

Belin, T. R., and D. B. Rubin. 1995. "A Method for Calibrating False- Match Rates in Record Linkage." *Journal of the American Statistical Association* 90, 694-707.

Benjelloun, O., and others. 2005. "Swoosh: A Generic Approach to Entity Resolution." Stanford University technical report, March 2005.

Bentley, J.L., and R. A. Sedgewick. 1997. "Fast Algorithms for Searching and Sorting Strings." *Proceedings of the Eighth ACM-SIAM Symposium on Discrete Algorithms*.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland. 1975. *Discrete Multivariate Analysis.* Cambridge: MIT Press.

Bilenko, M., and R. J. Mooney. 2003. “Adaptive Duplicate Detection Using Learnable String Similarity Metrics.” *Proceedings of ACM Conference on Knowledge Discovery and Data Mining.* Washington, ACM.

———. 2003. “On Evaluation and Training-Set Construction for Duplicate Detection.” *Proceedings of the ACM Workshop on Data Cleaning, Record Linkage and Object Identification.* Washington: ACM.

Bilenko, M., and others. 2003. “Adaptive Name Matching in Information Integration.” *IEEE Intelligent Systems* 18 (50): 16-23.

Bilke, A., and F. Naumann. 2005. “Schema Matching Using Duplicates.*” IEEE International Conference on Data Engineering*, 00: 69-80.

Borkar, V., K. Deshmukh, and S. Sarawagi. 2001. “Automatic Segmentation of Text into Structured Records.” (Washington: ACM, SIGMOD).

Borthwick, A. 2002. “MEDD 2.0.” Paper presented at conference New York, February 2002, available at www.choicemaker.com.

Broadbent, K., and W. Iwig. 1999. “Record Linkage at NASS using AutoMatch.” Paper presented at the Federal Committee on Statistical Methodology (FCSM) research conference, November 15-17, 1999, Arlington, VA. Available at www.fcsm.gov/99papers/broadbent.pdf.

Chaudhuri, S and others. 2003. “Robust and Efficient Match for On-Line Data Cleaning.” Paper presented at the ACM SIGMOD 2003 conference, San Diego, California June 9-12, 2003, .

Chaudhuri, S., V. Ganti, and R. Motwani. 2005. “Robust Identification of Fuzzy Duplicates.” IEEE International Conference on Data Engineering, 00: 865-876.

Christen, P. T. Churches, and J. X. Zhu. 2002. “Probabilistic Name and Address Cleaning and Standardization.” Paper presented at the Australian Data Mining Workshop, November, available at http://datamining.anu.edu.au/projects/linkage.html.

Churches, T. and others. 2002. “Preparation of Name and Address Data for Record Linkage Using Hidden Markov Models.” BioMed Central Medical Informatics and Decision Making 2 (9), available at http://www.biomedcentral.com/1472-6947/2/9/.

Cohen, W. W., P. Ravikumar, and S. E. Fienberg. 2003. “A Comparison of String Metrics for Matching Names and Addresses.” *International Joint Conference on Artificial Intel*ligence, *Proceedings of the Workshop on Information Integration on the Web*, Acapulco, Mexico, August.

———. 2003. “A Comparison of String Distance Metrics for Name-Matching Tasks.” *Proceedings of the ACM Workshop on Data Cleaning, Record Linkage and Object Identification.* Washington: ACM.

Cohen, W. W., and J. Richman. 2002. “Learning to Match and Cluster Entity Names.” Paper presented at the ACM Knowledge Discovery and Data Mining Conference, July 23 - 26, **2002**. Edmonton, Alberta, Canada .

Cohen, W. W., and S. Sarawagi. 2004. “Exploiting Dictionaries in Named Entity Extraction: Combining Semi-Markov Extraction Processes and Data Integration Methods.” Paper presented at the ACM Knowledge Discovery and Data Mining Conference, August 21-24, **2005** Chicago, IL.

Copas, J. R., and F. J. Hilton. 1990. "Record Linkage: Statistical Models for Matching Computer Records." *Journal of the Royal Statistical Society* A, 153: 287-320.

Cozman, F. G., I. Cohen, and M. C. Circio. 2003. “Semi-Supervised Learning of Mixture Models.” In T. Fawcett and N. Mishra, eds., *Proceedings of the Twentieth International Conference on Machine Learning* Cambridge, MA: American Association for Artificial Intelligence Press.

Dong, X., A. Halevy, and J. Madhavan. 2005. “Reference Reconciliation in Complex Information Spaces.” *Proceedings of the ACM SIGMOD Conference*. Washington: ACM.

DeGuire, Y. 1988. "Postal Address Analysis." *Survey Methodology* 14: 317-325.

Della Pietra, S., V. Della Pietra, and J. Lafferty. 1997. “Inducing Features of Random Fields.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19: 380-393.

Deming, W. E., and G. J. Gleser. 1959. "On the Problem of Matching Lists by Samples." *Journal of the American Statistical Association* 54: 403-415.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society*, B 39: 1-38.

Denis, F., and others. 2003. “Text Classification and Co-Training from Positive and Unlabeled Examples.” Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining, International Conference on Machine Learning.

Dhillon, I. S., Mallela, S., and Kumar, R. 2003. “A Divisive Information-Theoretic Feature Clustering Algorithm for Text Classification.” *Journal of Machine Learning Research* 3: 1265-1287.

Do, H. H., and E. Rahm. 2002. “COMA – A System for Flexible Combination of Schema Matching Approaches.” *Very Large Data Bases* 20: 610-621.

Elfekey, M., V. Vassilios, and A. Elmagarmid. 2002. “TAILOR: A Record Linkage Toolbox.” *IEEE International Conference on Data Engineering* 2002: 17-28.

Faloutsos, C., and K. I. Lin. 1995. “FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets.” *Proceedings of the ACM SIGMOD Conference.* New York: ACM.

Fellegi, I. P., and A. B. Sunter. 1969. "A Theory for Record Linkage." *Journal of the American Statistical Association* 64: 1183-1210.

Ferragina, P., and R. Grossi. 1999. “The String B-Tree: A New Data Structure for String Search in External Memory and Its Applications.“ *Journal of the Association of Computing Machinery* 46: 236-280.

Friedman, J. 2001. “Greedy Function Approximation: A Gradient Boosting Machine.” *Annals of Statistics* 29 (5): 1389-1432.

Gill, L. 1999. “OX-LINK: The Oxford Medical Record Linkage System.” In *Record Linkage Techniques 1997*. Washington: National Academy Press.

Getoor, L., and others. 2003. “Learning Probabilistic Models for Link Structure.” *Journal of Machine Learning Research* 3: 679-707.

Guha, S., and others. 2004. “Merging the Results of Approximate Match Operations.” *Proceedings of the 30th VLDB Conference*.

Hall, P. A. V., and G. R. Dowling. 1980. "Approximate String Comparison." *Association of Computing Machinery, Computing Surveys* 12: 381-402.

Hastie, T., R. Tibshirani, and J. Friedman. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer.

Hjaltson, G., and H. Samet. 2003. “Index-Driven Similarity Search in Metric Spaces.” *ACM Transactions on Database Systems* 28 (4): 517-580.

Jaro, M. A. 1989. "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida." *Journal of the American Statistical Association* 89: 414-420.

Jin, L., C. Li, and S. Mehrortra. 2002. “Efficient String Similarity Joins in Large Data Sets.” UCI Technical Report, Feb. 2002, available at www.ics.uci.edu/~chenli/pub/strjoin.pdf.

———. 2003. “Efficient Record Linkage in Large Data Sets.” Eight International Conference for Database Systems for Advance Applications (DASFAA 2003), 26-28 March, 2003, Kyoto, Japan, available at www.ics.uci.edu/~chenli/pub/dasfaa03.pdf .

Kim, J. J., and W. E. Winkler. 1995. “Masking Microdata Files.”, *Proceedings of the Section on Survey Research Methods.* Washington: American Statistical Association.

———. 2001. “Multiplicative Noise for Masking Continuous Data.” *American Statistical Association*, *Proceedings of the Section on Survey Research Methods.* Washington: ASA, CD-ROM.

Koller, D., and A. Pfeffer. 1998. “Probabilistic Frame-Based Systems.” *Proceedings of the Fifteenth National Conference on Artificial Intelligence.* Madison, WI: AAAI Press / The MIT Press, 580-587.

Koudas, N., A. Marathe, and D. Srivastava. 2004. “Flexible String Matching Against Large Databases in Practice.” *Operations: Proceedings of the 30th VLDB Conference*.

Lafferty, J., A. McCallum, and F. Pereira. 2001. “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.” *Proceedings of the International Conference on Machine Learning*. Cambridge, MA: American Association for Artificial Intelligence Press.

Larsen, M. 1999. “Multiple Imputation Analysis of Records Linked Using Mixture Models.” *Statistical Society of Canada Proceedings of the Survey Methods Section*. Montreal.

Lahiri, P., and M. D. Larsen. 2000. “Model-Based Analysis of Records Linked Using Mixture Models.” *Proceedings of the Section on Survey Research Methods.* Washington: American Statistical Association.

Lahiri, P. A., and M.D. Larsen. 2004. “Regression Analysis with Linked Data.” *Journal of the American Statistical Association*, 85, to appear.

Larsen, M. D., and D. B. Rubin. 2001. “Alternative Automated Record Linkage Using Mixture Models.” *Journal of the American Statistical Association* 79: 32-41.

Lu, Q., and L. Getoor. 2003. “Link-based Classification.” In T. Fawcett and N. Mishra, eds., *Proceedings of the Twentieth International Conference on Machine Learning.* Cambridge, MA: American Association for Artificial Intelligence Press..

Malin, B., L. Sweeney, and E. Newton. 2003. “Trail Re-Identification: Learning Who You Are From Where You Have Been.” Workshop on Privacy in Data, Carnegie-Mellon University, March.

McCallum, A., K. Nigam, and L. H. Unger. 2000. “Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching.” In *Knowledge Discovery and Data Mining.* New York: ACM Press.

McCallum, A., and B. Wellner. 2003. “Object Consolidation by Graph Partitioning with a Conditionally-Trained Distance Metric.” *Proceedings of the ACM Workshop on Data Cleaning, Record Linkage and Object Identification.* Washington: ACM.

McGovern, A., and D. Jensen. 2003. “Semi-Supervised Learning of Mixture Models.” In T. Fawcett and N. Mishra, eds., *Proceedings of the Twentieth International Conference on Machine Learning.* Cambridge, MA: American Association for Artificial Intelligence Press.

Meng, X., and D. B. Rubin. 1991. "Using EM to Obtain Asymptotic Variance-Covariance Matrices: The SEM Algorithm." *Journal of the American Statistical Association* 86: 899-909.

———. 1993. "Maximum Likelihood Via the ECM Algorithm: A General Framework." *Biometrika* 80: 267-278.

Michalowski, M., S. Thakkar, and C. A. Knoblock. 2003. “Exploiting Secondary Sources for Object Consolidation.” *Proceedings of the ACM Workshop on Data Cleaning, Record Linkage and Object Identification.* Washington: ACM.

———. 2004. “Exploiting Secondary Sources for Unsupervised Record Linkage.” *Proceedings of the 30th VLDB Conference.* Toronto, Canada: Morgan Kaufmann Publishers Inc.

Navarro, G. 2001. “A Guided Tour of Approximate String Matching.” *Association of Computing Machinery Computing Surveys* 33: 31-88.

Neiling, M., and S. Jurk. 2003. “The Object Identification Framework.” *Proceedings of the ACM Workshop on Data Cleaning, Record Linkage and Object Identification.* Washington: ACM.

Neter, J., E. S. Maynes, and R. Ramanathan. 1965. "The Effect of Mismatching on the Measurement of Response Errors." *Journal of the American Statistical Association* 60: 1005-1027.

Newcombe, H. B. 1988. *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration*, *and Business.* Oxford: Oxford University Press.

Newcombe, H. B., and others. 1959. "Automatic Linkage of Vital Records." *Science* 130: 954-959.

Newcombe, H.B., and J. M. Kennedy. 1962. "Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information." *Communications of the Association for Computing Machinery* 5: 563-567.

Newcombe, H. B., and M. E. Smith. 1975. “Methods for Computer Linkage of Hospital Admission-Separation Records into Cumulative Health Histories.” *Methods of Information in Medicine* 14 (3): 118-125.

Norén, G. N., R. Orre, and A. Bate. 2005. “A Hit-Miss Model for Duplicate Detection in the WHO Drug Safety Database.” *Proceedings of the ACM KDD Conference*. Washington: ACM.

Pasula, H., and others. 2003. “Identity Uncertainty and Citation Matching.” *Neural Information Processing Systems.* NIPS 15. MIT Press, Cambridge, MA 2003.

Pasula, H., and S. Russell. 2001. “Approximate Inference for First-Order Probabilistic Languages,” *Proceedings of the International Joint Conference on Artificial Intelligence*.

Pasula, H., and others. 1999. “Tracking Many Objects with Many Sensors.” *Proceedings of the Joint International Conference on Artificial Intelligence*.

Pollock, J., and A. Zamora. 1984. "Automatic Spelling Correction in Scientific and Scholarly Text." *Communications of the ACM* 27: 358-368.

Porter, E. H., and W. E. Winkler. 1999. “Approximate String Comparison and Its Effect in an Advanced Record Linkage System.” In Alvey and Jamerson, eds., *Record Linkage Techniques – 1997.* Washington: National Academy Press.

Rahm, E., and H. H. Do. 2000. “Data Cleaning: Problems and Current Approaches.” *IEEE Bulletin on Data Engineering* 23 (4): 3-13.

Ravikumar, P., and W. W. Cohen. 2004. “A Hierarchical Graphical Model for Record Linkage.” *Proceedings of the Conference on Uncertainty in Artificial Intelligence*.

Ristad, E. S., and P. Yianilos. 1998. “Learning String-Edit Distance.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* : 20 (2) 522-531.

Russell, S. 2001. “Identity Uncertainty.” *Proceedings of IFSA-01*. MIT Press, Cambridge, MA.

Sarawagi, S., and A. Bhamidipaty. 2002. “Interactive Deduplication Using Active Learning.” *Very Large Data Bases* 2002: 269-278.

Sarawagi, S., S. Chakrabarti, and S. Godbole. 2003. “Cross-Training: Learning Probabilistic Mappings between Topics.” In L. Getoor, ed., *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* Washington, DC, USA, August 24 – 27. Washington: ACM Press.

Scannipieco, M. 2003. “DAQUINCIS: Exchanging and Improving Data Quality in Cooperative Information Systems.” Ph.D. dissertation, University of Rome, La Sapienza.

Scheuren, F. 1980. “Methods of Estimation for the 1973 Exact Match Study.” *Studies from Interagency Data Linkages*. Report No. 101. Washington: U.S. Social Security Administration.

Scheuren, F., and W. E. Winkler. 1993. "Regression Analysis of Data Files That Are Computer Matched." *Survey Methodology* 19: 39-58.

———. 1997. "Regression Analysis of Data Files That Are Computer Matched, II." *Survey Methodology* 23: 157-165.

Sekar, C. C., and W. E. Deming. 1949. "On a Method of Estimating Birth and Death Rates and the Extent of Registration." *Journal of the American Statistical Association* 44: 101-115.

Taskar, B., P. Abdeel, and D. Koller. 2002. “Discriminative Probabilistic Models for Relational Data.” In Proceedings. of Eighteenth Conference On Uncertainty in Artificial Intelligence (UAI02), Edmonton, Canada, 2003.

Taskar, B., E. Segal, and D. Koller. 2001. “Probabilistic Classification and Clustering in Relational Data.” In Bernhard Nebel, editor, Proceeding of IJCAI-01, *17th International Joint Conference on Artificial Intelligence*, pages 870--878, Seattle, US, 2001.

Taskar, B., and others. 2003. “Link Prediction in Relational Data.” From the *Neural Information Processing Systems Conference 2003.* Published in Sebastian Thrun, and others. Eds. *Advances in Neural Information Processing Systems 16*. Cambridge, MA: MIT Press. Also available at http://books.nips.cc/nips16.html.

Taskar, B., M. F. Wong, and D. Koller. 2003. “Learning on Test Data: Leveraging ‘Unseen’ Features.” *Proceedings of the Twentieth International Conference on Machine Learning.* Cambridge, MA: American Association for Artificial Intelligence Press.

Thibaudeau, Y. 1989. "Fitting Log-Linear Models When Some Dichotomous Variables are Unobservable." *Proceedings of the Section on Statistical Computing.* Washington: American Statistical Association..

———. 1993. "The Discrimination Power of Dependency Structures in Record Linkage." *Survey Methodology* 19: 31-38.

Titterington, D. M., A. Smith, and U. E. Makov. 1988. *Statistical Analysis of Finite Mixture Distributions.* New York: Wiley.

Torra, V. 2000. “Re-Identifying Individuals Using OWA Operators.” *Proceedings of the Sixth Conference on Soft Computing*. New York: Springer-Verlag, Inc.

———. 2004. “OWA Operators in Data Mining: Modeling and Re-Identification.” *IEEE Transactions on Fuzzy Systems*. Piscataway, NJ: The Institute of Electrical and Electronics Engineers, Inc.

Vapnik, V. 2000. *The Nature of Statistical Learning Theory,* 2nd edition. Berlin: Springer.

Wang, S., and others. 2003. “Learning Mixture Models with the Latent Maximum Entropy Principal.” In T. Fawcett and N. Mishra, eds., *Proceedings of the Twentieth International Conference on Machine Learning*. Cambridge, MA: American Association for Artificial Intelligence Press.

Wei, J. 2004. “Markov Edit Distance.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26 (3): 311-321.

Winkler, W. E. 1988. "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage." *Proceedings of the Section on Survey Research Methods.* Washington: American Statistical Association.

———. 1989. "Near Automatic Weight Computation in the Fellegi-Sunter Model of Record Linkage." *Proceedings of the Fifth Census Bureau Annual Research Conference*. Washington, DC: Government Printing Office.

———. 1989. "Methods for Adjusting for Lack of Independence in an Application of the Fellegi-Sunter Model of Record Linkage." *Survey Methodology* 15: 101-117.

———. 1989. "Frequency-based Matching in the Fellegi-Sunter Model of Record Linkage." *Proceedings of the Section on Survey Research Methods.* Washington: American Statistical Association..

———. 1990. "Documentation of Record-Linkage Software." Unpublished report. Washington: Statistical Research Division, U.S. Bureau of the Census.

———. 1990. "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage." *Proceedings of the Section on Survey Research Methods.* Washington: *American Statistical Association*.

———. 1990. “On Dykstra’s Iterative Fitting Procedure.” *Annals of Probability* 18: 1410-1415.

———. 1993. "Business Name Parsing and Standardization Software." Unpublished report. Washington: Statistical Research Division, U.S. Bureau of the Census.

———. 1993. "Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage." *Proceedings of the Section on Survey Research Methods.* Washington: American Statistical Association.

———. 1994. "Advanced Methods for Record Linkage." *Proceedings of the Survey Research Methods Section.* Washington: *American Statistical Association* (longer version report 94/05 available at www.census.gov/srd/www/byyear.html.

———. 1995. "Matching and Record Linkage." In B. G. Cox and others, eds., *Business Survey Methods.* New York: Wiley, also available at http://www.fcsm.gov/working-papers/wwinkler.pdf.

———. 1997. "Producing Public Use Microdata That are Analytically Valid and Confidential." *Proceedings of the Section on Survey Research Methods.* Washington: American Statistical Association.

———. 1998. “Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata.” *Research in Official Statistics* 1: 87-104.

———. 1999. "The State of Record Linkage and Current Research Problems." *Proceedings of the Survey Methods Section.* Montreal: Statistical Society of Canada (longer version available at www.census.gov/srd/www/byyear.html).

———. 1999. “Issues with Linking Files and Performing Analyses on the Merged Files.” *Proceedings of the Sections on Government Statistics and Social Statistics.* Washington: American Statistical Association.

———. 1999. "Record Linkage Software and Methods for Administrative Lists." *Proceedings of the Exchange of Technology and Know-How '99.* Eurostat, also available at www.census.gov/srd/www/byyear.html.

———. 2000. “Machine Learning, Information Retrieval, and Record Linkage.”*Proceedings of the Survey Research Methods Section.* Washington: American Statistical Association, also available at www.niss.org/affiliates/dqworkshop/papers/winkler.pdf.

———. 2001. “The Quality of Very Large Databases.” *Proceedings of Quality in Official Statistics 2001.* CD-ROM. May 15-17, 2001, Stockholm, Sweden, 2001 Also available at www.census.gov/srd/www/byyear.html as report rr01/04.

———. 2001. “Record Linkage.” In A. H. El-Shaarawi and W. W. Piegorsch, eds., *Encyclopedia on Environmetrics.* New York: Wiley.

———. 2002. “Record Linkage and Bayesian Networks.” *Proceedings of the Survey Research Methods Section.* Washington: American Statistical Association, CD-ROM. Also at www.census.gov/srd/www/byyear.html).

———. 2003. “Methods for Evaluating and Creating Data Quality.” *Proceedings of the ICDT Workshop on Cooperative Information Systems* (Siena, IT. January 2003). Longer version in *Information Systems* 29 (7) (2004): 531-550.

———. 2003. “Data Cleaning Methods.” *Proceedings of the ACM Workshop on Data Cleaning, Record Linkage and Object Identification.* Washington: ACM.

———. 2004. “Re-identification Methods for Masked Microdata.” In J. Domingo-Ferrer and V. Torra, eds., *Privacy in Statistical Databases 2004.* New York: Springer, available at www.census.gov/srd/papers/pdf/rrs2004-03.pdf.

———. 2004. “Masking and Re-identification Methods for Public use Microdata: Overview and Research Problems.” In J. Domingo-Ferrer and V. Torra, eds., *Privacy in Statistical Databases 2004.* New York: Springer, available at www.census.gov/srd/papers/pdf/rrs2004-06.pdf.

———. 2004. “Record Linkage: Overview of Recent Developments and Applications.” In S. Biffignandi, ed., *Combining Data from Different Sources – Applications of Record Linkage Methodology and Estimation Using Administrative Data.* Rome: ISTAT.

———. 2004. “Approximate String Comparator Search Strategies for Very Large Administrative Lists.” *Proceedings of the Section on Survey Research Methods.* Washington: American Statistical Association, available at www.census.gov/srd/www/byyear.html.

———. 2005. “Overview of Record Linkage and Current Research Directions.” Washington: U.S. Bureau of the Census, Statistical Research Division Report, available at www.census.gov/srd/www/byyear.html.

———. 2005. “Data Quality in Data Warehouses.” In John Wang, ed., *Encyclopedia of Data Warehousing and Data Mining*. Hershey, PA: Idea Group Publishing, Inc.

———. (2005c), “Methods and Analyses for Determining Quality,” to appear.

Winkler, W. E., and F. Scheuren. 1991. "How Computer Matching Error Effects Regression Analysis: Exploratory and Confirmatory Analysis." Technical Report. Washington: U.S. Bureau of the Census, Statistical Research Division.

———. 1995. "Linking Data to Create Information," *Proceedings of Symposium 95: From Data to Information - Methods and Systems.* Montreal: Statistics Canada.

———. 1996. "Recursive Analysis of Linked Data Files." *Proceedings of the 1996 Census Bureau Annual Research Conference.* Washington, DC: Government Printing Office.

Winkler, W. E., and Y. Thibaudeau. 1991. "An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Census." Technical report. Washington: U.S. Bureau of the Census, Statistical Research Division, available at www.census.gov/srd/www/byyear.html.

Yancey, W.E. 2000. “Frequency-Dependent Probability Measures for Record Linkage.” *Proceedings of the Section on Survey Research Methods.* Washington: American Statistical Association, available at www.census.gov/srd/www/byyear.html.

———. 2002. “Improving EM Parameter Estimates for Record Linkage Parameters.” *Proceedings of the Survey Research Methods Section.* Washington; American Statistical Association, CD-ROM. Available as Report RRS 2004/01at www.census.gov/srd/www/byyear.html.

———. 2003. “An Adaptive String Comparator for Record Linkage.” *Proceedings of the Survey Research Methods Section.* Washington: American Statistical Association, CD-ROM. Also available as Report RRS 2004/02 at www.census.gov/srd/www/byyear.html.

———. 2005. “Evaluating String Comparator Performance for Record Linkage.” Research Report RRS 2005/05. Washington: U.S. Census Bureau, available at www.census.gov/srd/www/byyear.html.

Yancey, W.E., and W. E. Winkler. 2003. “BigMatch Software.” Computer System, Documentation. Washington: U. S. Census Bureau, available at www.census.gov/srd/www/byyear.html.

Yancey, W.E., W.E. Winkler, and R. H. Creecy. 2002.“Disclosure Risk Assessment in Perturbative Microdata Protection.” In J. Domingo-Ferrer, ed., *Inference Control in Statistical Databases.* New York: Springer, available at www.census.gov/srd/papers/pdf/rrs2002-01.pdf.

THE BROOKINGS INSTITUTION

1775 Massachusetts Avenue, NW • Washington, DC 20036-2188
Tel: 202-797-6000 • Fax: 202-797-6004
www.brookings.edu

METROPOLITAN POLICY PROGRAM

DIRECT: 202-797-6139 • FAX/DIRECT: 202-797-2965
www.brookings.edu/metro