

Confidentiality and Disclosure Limitation



Stephen E. Fienberg

Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

Glossary

confidentiality Broadly, a quality or condition accorded to statistical information as an obligation not to transmit that information to an unauthorized party.

contingency table A cross-classified table of counts according to two or more categorical variables.

data masking The disclosure limitation process of transforming a data set when there is a specific functional relationship (possibly stochastic) between the masked values and the original data.

disclosure The inappropriate attribution of information to a data provider, whether it be an individual or organization.

disclosure limitation The broad array of methods used to protect confidentiality of statistical data.

perturbation An approach to data masking in which the transformation involves random perturbations of the original data, either through the addition of noise or via some form of restricted randomization.

privacy In the context of data, usually the right of individuals to control the dissemination of information about themselves.

Confidentiality and privacy are widely conceived of as essential components of the collection and dissemination of social science data. But providing access to such data should also be a goal of social science researchers. Thus, researchers should attempt to release the maximal amount of information without undue risk of disclosure of individual information. Assessing this trade-off is inherently a statistical issue, as is the development of methods to limit disclosure risk. This article addresses some aspects of confidentiality and privacy as they relate to social science research and describes some basic disclosure

limitation approaches to protect confidentiality. It also outlines some of the evolving principles that are guiding the development of statistical methodology in this area.

Introduction and Themes

Social science data come in a wide variety of forms, at least some of which have been gathered originally for other purposes. Most of these databases have been assembled with carefully secured consent and cooperation of the respondents, often with pledges to keep the data confidential and to allow their use for statistical purposes only. The general public disquiet regarding privacy, spurred on by the privacy threats associated with Internet commerce and unauthorized access to large commercial databases (e.g., those maintained by banks and credit agencies), has heightened concerns about confidentiality and privacy in the social sciences and in government statistical agencies. In the universities, social science and, in particular, survey data have come under increased scrutiny by institutional review boards, both regarding pledges of confidentiality and the means that researchers use to ensure them. But social scientists and government statistical agencies also have an obligation to share data with others for replication and secondary analysis. Thus, researchers need to understand how to release the maximal amount of information without undue risk of disclosure of individual information.

For many years, confidentiality and disclosure limitation were relegated to the nonstatistical part of large-scale data collection efforts; as a consequence, the methods used to address the issue of privacy were often *ad hoc* and conservative, directed more at protection and less at the usability of the acquired data. More recently, statisticians have convinced others that any release of statistical data produces a disclosure in that it increases the

probability of identification of some individual in the relevant population. From this now widely recognized perspective, the goal of the preservation of promises of confidentiality cannot be absolute, but rather should be aimed, of necessity, at the limitation of disclosure risk rather than at its elimination. Assessing the trade-off between confidentiality and data access is inherently a statistical issue, as is the development of methods to limit disclosure risk. That is, formulation of the problem is statistical, based on inputs from both the data providers and the users, regarding both risk and utility.

The article covers some basic definitions of confidentiality and disclosure, the ethical themes associated with confidentiality and privacy, and the timing of release of restricted data to achieve confidentiality objectives, and albeit briefly, when release restricted data is required to achieve confidentiality or when simply restricting access is a necessity. A case is made for unlimited access to restricted data as an approach to limit disclosure risk, but not so much as to impair the vast majority of potential research uses of the data.

In recent years, many researchers have argued that the trade-off between protecting confidentiality (i.e., avoiding disclosure) and optimizing data access to others has become more complex, as both technological advances and public perceptions have not altered in an information age, but also that statistical disclosure techniques have kept pace with these changes. There is a brief introduction later in the article to some current methods in use for data disclosure limitation and statistical principles that underlie them. This article concludes with an overview of disclosure limitation methodology principles and a discussion of ethical issues and confidentiality concerns raised by new forms of statistical data.

What Is Meant by Confidentiality and Disclosure?

Confidentiality refers broadly to a quality or condition accorded to statistical information as an obligation not to transmit that information to an unauthorized party. It has meaning only when a data collector, e.g., a university researcher or a government statistical agency, can deliver on its promise to the data provider or respondent. Confidentiality can be accorded to both individuals and organizations; for the individual, it is rooted in the right to privacy (i.e., the right of individuals to control the dissemination of information about themselves, whereas for establishments and organizations, there are more limited rights to protection, e.g., in connection with commercial secrets.

Disclosure relates to inappropriate attribution of information to a data provider, whether to an individual or organization. There are basically two types of disclosure, identity and attribute. An identity disclosure occurs if the data provider is identifiable from the data release. An attribute disclosure occurs when the released data make it possible to infer the characteristics of an individual data provider more accurately than would have otherwise been possible. The usual way to achieve attribute disclosure is through identity disclosure; an individual is first identified through some combination of variables and then there is attribute disclosure of values of other variables included in the released data. However, attribute disclosure may occur without an identification. The example of union plumbers in Chicago has been used to elucidate this: for the plumbers, who all earn the same wage, attribute disclosure occurs when the Department of Labor releases the average wage of plumbers in Chicago as an entry in a table.

Inferential disclosure is basically a probabilistic notion; the definition can be interpreted as referring to some relevant likelihood ratio associated with the probability of identifying an attribute for a data provider. Because almost any data release can be expected to increase the likelihood associated with some characteristic for some data provider, the only way data protection can be guaranteed is to release no data at all. It is for this reason that the methods used to protect confidentiality are referred to in the statistical literature as disclosure limitation methods or statistical disclosure control, rather than disclosure prevention methods.

Clearly, one must remove names, addresses, telephone numbers, and other direct personal identifiers from databases to preserve confidentiality, but this is not sufficient. Residual data that are especially vulnerable to disclosure threats include (1) geographic detail, (2) longitudinal panel information, and (3) extreme values (e.g., on income). Population data are clearly more vulnerable than are sample data, and “key variables” that are also available in other databases accessible to an intruder pose special risks. Statistical organizations have traditionally focused on the issue of identity disclosure and thus refuse to report information in which individual respondents or data providers can be identified. This occurs, for example, when a provider is unique in the population for the characteristics under study and is directly identifiable in the database to be released. But such uniqueness and subsequent identity disclosure may not, of course, “reveal” any information other than that the respondent provided data as part of the study. In this sense, identity disclosure may be only a technical violation of a promise of confidentiality, but not a violation of the spirit of such a promise. Thus, uniqueness only raises the issue of “possible” confidentiality problems due to disclosure.

The foregoing discussion implicitly introduces the notion of harm, which is not the same as a breach of

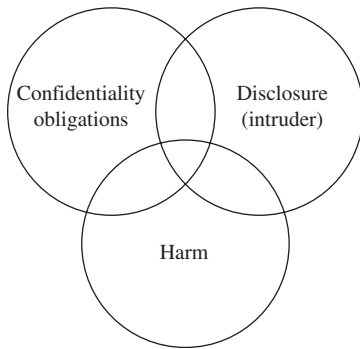


Figure 1 Relationship between confidentiality, disclosure, and harm. Adapted from Fienberg (2001).

confidentiality. As Fig. 1 depicts in a schematic fashion, not all data disclosures breach promises of confidentiality to respondents. For example, it is possible for a pledge of confidentiality to be technically violated, but for there to be no harm to the data provider because the information is “generally known” to the public at large. In this case, some would argue that additional data protection is not really required. Conversely, if an intruder attempts to match records from an external file to another file subject, and to a pledge of confidentiality, but makes an “incorrect” match, then there is no breach of confidentiality, but there is the possibility of harm if the intruder acts as though the match were correct. Further, information on individuals (or organizations) in a release of sample statistical data may well increase the information about characteristics of individuals (or organizations) not in the sample. This produces an inferential disclosure for such individuals (or organizations), causing them possible harm, even though there is no confidentiality obligation for those so harmed. For example, suppose it is possible to infer with high probability that, for a given set of variables to be released, there is a unique individual in the population corresponding to a zero cell in a sample cross-classification of these variables. Then there is a high probability of disclosure of that unique individual, even though he or she is not present in the sample of data released.

Some people believe that the way to assure confidentiality and prevent disclosures is allow participants/respondents to participate in a study anonymously. Except for extraordinary circumstances, such a belief is misguided, because there is a key distinction between collecting information anonymously and ensuring that personal identifiers are not used inappropriately. For example, survey investigators typically need personal identifier information on participants/respondents to carry out nonresponse follow-up and to perform quality control checks on collected data by reinterviewing respondents, even if such identifiers are not released to other users of the data. Moreover, as has already been noted, the simple

removal of personal identifiers is not sufficient to prevent disclosures.

Yet, there remain circumstances in which the disclosure of information provided by respondents under a pledge of confidentiality can produce substantial harm, in terms of personal reputations or even in a monetary sense, to respondents, their families, and others with whom they have personal relationships. For example, in the pilot surveys for the National Household Seroprevalence Survey, the National Center for Health Statistics moved to make responses during the data collection phase of the study truly “anonymous” because of the potential for harm resulting from the “release” of information either that the respondent tested positive for the human immunodeficiency virus or engaged in high-risk behavior. But such efforts still could not guarantee that an intruder could not identify someone in the survey database. This example also raises the interesting question about the applicability of confidentiality provisions after an individual’s death, in part because of the potential harm that might result to others. Several statistical agencies explicitly treat the identification of a deceased individual as a violation of confidentiality, even when there is no legal requirement to do so.

Thus there is a spectrum of types and severity of harm that might result from both nondisclosures and disclosures, and these must all be considered in some form when one develops a survey instrument, crafts a statement regarding confidentiality, or prepares a database for possible release. The present discussion restricts attention primarily to settings involving the possibility of either inferential or identity disclosure and the attendant potential harm to the data providers.

Restricted Access versus Restricted Data

Social science data, especially those gathered as part of censuses and major sample surveys or with government funding, meet two key tests that are usually applied to public goods: jointness of consumption (consumption by one person does not diminish their availability to others) and benefit to the social science enterprise and thus the nation as a whole (e.g., social science data are used to inform public policy). The only issue, then, is whether there is nonexclusivity, i.e., whether it makes sense to provide these statistical data to some citizens and not to others. If it is possible to provide access to all or virtually all, e.g., via the Internet and the World Wide Web, then the costs of providing the data to all are often less than the costs of restricting access. There are other perhaps hidden costs, however, that result from expanded use to those who produce the data. Several reports have

described the costs and benefits of data sharing and data sharing in the context of confidentiality. The principal U.S. research funding agencies, the National Science Foundation and the National Institutes of Health, now require data sharing as part of the contractual arrangements for grants. These policies continue a long tradition of broad access to survey research data through archives, such as the one in the United States operated by the Inter-University Consortium for Political and Social Research.

People adopt two different philosophies with regard to the preservation of confidentiality associated with individual-level data: (1) restricted or limited information, with restrictions on the amount or format of the data released, and (2) restricted or limited access, with restrictions on the access to the information. If social science data are truly a public good, then restricted access is justifiable only in extreme situations, when the confidentiality of data in the possession of a researcher or statistical agency (in the case of government data) cannot be protected through some form of restriction on the information released.

In many countries, the traditional arrangement for data access has been limited in both senses described here, with only highly aggregated data and summary statistics released for public consumption. In the United States, there have been active policies encouraging the release of microdata from censuses and samples going back to the 1960 decennial census. Such data are now publicly available in a variety of forms, including on-line data archives. Some countries have attempted to follow suit, but others have development approaches based on limited access. In fact, many have argued that, even if cost is not an issue, access should be restricted as a means for ensuring confidentiality, especially in the context of establishment data. For example, the U.S. Bureau of the Census has now established several data centers, including one in Pittsburgh (at Carnegie Mellon University), through which restricted access to confidential Census Bureau data sets can be arranged. The process of gaining access to data in such centers involves an examination of the research credentials of those wishing to do so. In addition, these centers employ a mechanism for controlling the physical access to confidential data files for those whose access is approved, and for the review of all materials that researchers wish to take from the centers and to publish. Just imagine the difficulty the researchers would have if they are accustomed to reporting residual plots and other information that allow for a partial reconstruction of the original data, at least for some variables, because restricted data centers typically do not allow users to take such information away.

A less extreme form of restricted access, and one more consonant with the notion of statistical data as a public good, has been adopted by a number of data archives. They often utilize a “terms-of-use agreement” via which

the secondary data analyst agrees to use the data for statistical research and/or teaching only, and to preserve confidentiality, even if data sets have already been edited using disclosure limitation methods. Some statistical agencies adopt related approaches of “licensing.” Restricting access to a “public good” produces bad public policy because it cannot work effectively. This is primarily because the gatekeepers for restricted data systems have little or no incentive to widen access or to allow research analysts the same freedom to work with a data set (and to share their results) as they are able to have with unrestricted access. And the gatekeepers can prevent access by those who may hold contrary views on either methods of statistical analyses or on policy issues that the data may inform. In this sense, the public good is better served by uncontrolled access to restricted data rather than by restricted access to data that may pose confidentiality concerns. This presumes, of course, that researchers are able to do an effective job of statistical disclosure limitation.

Methodology for Disclosure Limitation

Many disclosure limitation methods can be described under the broad rubric of disclosure-limiting masks, i.e., transformations of the data whereby there is a specific functional relationship (possibly stochastic) between the masked values and the original data. The basic idea of data masking involves thinking in terms of transformations, to transforming an $n \times p$ data matrix Z through pre- and postmultiplication, and the possible addition of noise, i.e.,

$$Z \rightarrow AZB + C, \quad (1)$$

where A is a matrix that operates on cases, B is a matrix that operates on variables, and C is a matrix that adds perturbations or noise. Matrix masking includes a wide variety of standard approaches to disclosure limitation:

- Adding noise
- Releasing a subset of observations (delete rows from Z)
- Cell suppression for cross-classifications
- Including simulated data (add rows to Z)
- Releasing a subset of variables (delete columns from Z)
- Switching selected column values for pairs of rows (data swapping)

Little was one of the first to describe likelihood-based methods for the statistical analysis of masked data. Even when a mask is applied to a data set, the possibilities of both

identity and attribute disclosure remain, although the risks may be substantially diminished.

It is possible to categorize most disclosure-limiting masks as suppressions (e.g., cell suppression), recodings (e.g., collapsing rows or columns, or swapping), samplings (e.g., releasing subsets), or simulations. Further, some masking methods alter the data in systematic ways, e.g., through aggregation or through cell suppression, and others do it through random perturbations, often subject to constraints for aggregates. Examples of perturbation methods are controlled random rounding, data swapping, and the postrandomization method (PRAM). One way to think about random perturbation methods is as a restricted simulation tool, and thus it is possible to link them to other types of simulation approaches that have recently been proposed.

Fienberg and colleagues have pursued this simulation strategy and presented a general approach to “simulating” from a constrained version of the cumulative empirical distribution function of the data. In the case when all of the variables are categorical, the cumulative distribution function is essentially the same as the counts in the resulting cross-classification or contingency table. As a consequence, this general simulation approach can be considered as equivalent to simulating from a constrained contingency table, e.g., given a specific set of marginal totals and replacing the original data by a randomly generated one drawn from the “exact” distribution of the contingency table under a log-linear model that includes “confidentiality-preserving” margins among its minimal sufficient statistics. If the simulated table is consistent with some more complex log-linear model, then this approach offers the prospect of simultaneously smoothing the original counts, offering room for model search and assessing goodness-of-fit, and providing disclosure limitation protection.

Rubin and others have asserted that the risk of identity disclosure can be eliminated by the use of synthetic data (using Bayesian methodology and multiple imputation techniques), because there is no direct function link between the original data and the released data. Or said another way, there is no confidentiality problem because all of the real individuals have been replaced with simulated ones. Raghunathan and colleagues describe the implementation of multiple imputation for disclosure limitation and a number of authors have now used variations on the approach, e.g., for longitudinally-linked individual and work history data. But with both simulation and multiple-imputation methodology, it is still possible that some simulated individuals may be virtually identical to original sample individuals in terms of their data values, or at least close enough that the possibility of both identity disclosure and attribute disclosure remains. Thus, it is still necessary to carry out checks for the possibility of unacceptable disclosure risk.

Another important feature of this statistical simulation approach is that information on the variability is directly accessible to the user. For example in the Fienberg *et al.* approach for categorical data, anyone can begin with the reported table and information about the margins that are held fixed, and then run the Diaconis–Sturmfels Monte Carlo Markov chain algorithm to regenerate the full distribution of all possible tables with those margins. This then allows the user to make inferences about the added variability in a modeling context. Similarly, multiple imputation can be used to get direct measure of variability associated with the posterior distribution of the quantities of interest. As a consequence, simulation and perturbation methods represent a major improvement from the perspective of access to data over cell suppression and data swapping. And they conform to a statistical principle of allowing the user of released data to apply standard statistical operations without being misled.

There has been considerable research on disclosure limitation methods for tabular data, especially in the form of multiway tables of counts (contingency tables). The most popular methods include collapsing categories (a form of aggregation) and a process known as cell suppression (developed by Larry Cox and others). Cell suppression systematically deletes the values in selected cells in the table. Though cell suppression methods have been very popular with the U.S. government statistical agencies and they are useful for tables with non-negative entries rather than simply counts, they also have major drawbacks. First, there are not yet good algorithms for the methodology associated with high-dimensional tables. But more importantly, the methodology systematically distorts for users the information about the cells in the table, and as a consequence, it makes it difficult for secondary users to draw correct statistical inferences about the relationships among the variables in the table.

A special example of collapsing involves summing over variables to produce marginal tables. Thus, instead of reporting the full multiway contingency table, one or more collapsed versions of it might be reported. The release of multiple sets of marginal totals has the virtue of allowing statistical inferences about the relationships among the variables in the original table using log-linear model methods. What is also intuitively clear from statistical theory is that, with multiple collapsed versions, there might be highly accurate information about the actual cell entries in the original table, and thus there will still be a need to investigate the possibility of disclosures. A number of researchers have been working on the problem of determining upper and lower bounds on the cells of a multiway table, given a set of margins, in part to address this problem, although other measures of risk may clearly be of interest. The problem of computing bounds is in one sense an old one (at least for two-way

tables), but it is also deeply linked to recent mathematical statistical developments and thus has generated a flurry of new research.

Consider a 2×2 table of counts $\{n_{ij}\}$ with given the marginal totals $\{n_{1+}, n_{2+}\}$ and $\{n_{+1}, n_{+2}\}$. The marginal constraints, i.e., that the counts in any row add to the corresponding one-way total, plus the fact that the counts must be nonnegative, imply bounds for the cell entries. Specifically, for the (i, j) cell,

$$\min\{n_{i+}, n_{+j}\} \geq n_{ij} \geq \max\{n_{i+} + n_{+j} - n, 0\}. \quad (2)$$

Bounds such as those in Eq. (2) usually are referred to as Fréchet bounds, after the French statistician M. Fréchet, but they were independently described by both Bonferroni and Hoeffding at about the same time in 1940. These bounds have been repeatedly rediscovered by a myriad of others. Fréchet bounds and their generalizations lie at the heart of a number of different approaches to disclosure limitation, including cell suppression, data swapping and other random perturbation methods, and controlled rounding.

Dobra and Fienberg have described multi-dimensional generalizations of the Fréchet bounds and explained some of the links between them and the modern statistical theory of log-linear models for the analysis of contingency tables. They analyzed a specific 6-way table and explained the extent to which easily computable bounds can be used to assess risk and data can be provided to users for model selection and assessing goodness-of-fit. Dobra and colleagues applied related methodology to a disclosure assessment of a 16-way contingency table drawn from the National Long Term Care Survey.

In the past decade, several special issues of statistical journals and edited volumes have highlighted the latest research on disclosure limitation methodology. Though many theoretical and empirical issues remain to be explored—for example, in connection with large sparse tables and longitudinal survey data—and many exciting research questions remain to be answered in this relatively new statistical literature on disclosure limitation research, the topic is receiving considerable attention and there is now a real prospect of improved disclosure limitation and increased data access in the not too distant future.

Conclusions and Further Issues

The focus here has been on the interplay between the issues of confidentiality and access to social science data. Disclosure limitation is an inherently statistical issue because the risk of disclosure cannot be eliminated unless access to the data is restricted. Complex relationships exist between promises of confidentiality to respondents in

surveys or to participants in studies and the nature of disclosure of information about those respondents. Because techniques for disclosure limitation are inherently statistical in nature, they must be evaluated using statistical tools for assessing the risk of harm to respondents. This article has outlined some of the current statistical methods used to limit disclosure, especially those representable in form of disclosure limitation masks, distinguishing among suppression, recoding, sampling, and simulation approaches, on the one hand, and systematic versus perturbational approaches on the other.

Among the principles that have been the focus of much of the recent effort in disclosure limitation methodology are usability, transparency, and duality. Usability is the extent to which the released data are free from systematic distortions that impair statistical methodology and inference. Transparency is the extent to which the methodology and practice of it provide direct or even implicit information on the bias and variability resulting from the application of a disclosure limitation mask. Duality is the extent to which the methods aim at both disclosure limitation and making the maximal amount of data available for analysis. The focus here in particular has been on how these principles fit with recent proposals for the release of simulated data, the release of marginals from multiway contingency tables, and the role of marginal bounds in evaluating the disclosure limitation possibilities.

As social scientists move to study biological correlates of social phenomena such as aging, they are beginning to incorporate direct measurements of health status based on tests, including those involving the drawing of blood samples. Technology and biological knowledge have advanced sufficiently that researchers now face the prospect of including in social science databases genetic sequencing information and other nonnumeric information, e.g., functional magnetic resonance imaging or full-body scan images. These new forms of data, in principle, can uniquely identify individuals. It is essential that social scientists begin thinking about how to handle the release of such data or how to limit their disclosure, through restriction of the data they make available to others. Such issues pose enormous methodological challenges for disclosure limitation research and for social science research data more broadly.

Acknowledgments

This work was supported in part by Grant No. EIA-9876619 from the U.S. National Science Foundation to the National Institute of Statistical Sciences and Grant No. R01-AG023141 from the National Institutes of Health.

See Also the Following Articles

Contingency Tables and Log-Linear Models • Statistical Disclosure Control

Further Reading

- Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge.
- Dobra, A., and Fienberg, S. E. (2000). Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *Proc. Natl. Acad. Sci.* **97**, 11885–11892.
- Dobra, A., and Fienberg, S. E. (2001). Bounds for cell entries in contingency tables induced by fixed marginal totals with applications to disclosure limitation. *Statist. J. UN, ECE* **18**, 363–371.
- Dobra, A., Erosheva, E., and Fienberg, S. E. (2003). Disclosure limitation methods based on bounds for large contingency tables with application to disability data. In *Proceedings of the Conference on New Frontiers of Statistical Data Mining* (H. Bozdogan, ed.), pp. 93–116. CRC Press, Boca Raton, Florida.
- Domingo-Ferrer, J. (ed.) (2002). Inference control in statistical databases from theory to practice. *Lecture Notes in Computer Science*, Vol. 2316. Springer-Verlag, Heidelberg.
- Doyle, P., Lane, J., Theeuwes, J., and Zayatz, L. (eds.) (2001). *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*. Elsevier, New York.
- Duncan, G. T. (2001). Confidentiality and statistical disclosure limitation. In *International Encyclopedia of the Social and Behavioral Sciences* (N. Smelser and P. Baltes, eds.), Vol. 4, pp. 2521–2525. Elsevier, New York.
- Fienberg, S. E. (2001). Statistical perspectives on confidentiality and data access in public health. *Statist. Med.* **20**, 1347–1356.
- Fienberg, S. E., Makov, U. E., and Steele, R. J. (1998). Disclosure limitation using perturbation and related methods for categorical data (with discussion). *J. Official Statist.* **14**, 485–511.
- Gouweleeuw, J. M., Kooiman, P., Willenborg, L. C. R. J., and Wolf, P. P. D. E. (1998). Post randomization for statistical disclosure control: Theory and implementation. *J. Official Statist.* **14**, 463–478.
- Lambert, D. (1993). Measures of disclosure risk and harm. *J. Official Statist.* **9**, 313–331.
- Little, R. J. A. (1993). Statistical analysis of masked data. *J. Official Statist.* **9**, 407–426.
- Raghunathan, T. E., Reiter, J., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *J. Official Statist.* **19**, 1–16.
- Trottini, M., and Fienberg, S. E. (2002). Modelling user uncertainty for disclosure risk and data utility. *Int. J. Uncertainty, Fuzziness Knowledge-Based Syst.* **10**, 511–528.
- Willenborg L., and De Waal, T. (2001). Elements of disclosure control. *Lecture Notes in Statistics*, Vol. 155. Springer-Verlag, New York.