A Disclosure Avoidance Research Agenda

Paul B. Massell

Center for Disclosure Avoidance Research, U.S. Census Bureau

Rm. 5K116A, 4600 Silver Hill Road, Washington, D.C. 20233, paul.b.massell@census.gov

Proceedings of the 2013 Federal Committee on Statistical Methodology (FCSM) Research Conference

Abstract

One of the striking facts about disclosure avoidance (also known by names such as 'confidentiality protection' or 'statistical disclosure control (SDC)') is the diversity of the protection methods. They vary greatly with the type of data product being protected, e.g., (frequency) count tables, magnitude data tables, microdata from a survey or a census or a statistical model. Some of the older methods for protecting count tables, e.g., collapsing of categories by which rows and columns are defined, can be learned by someone with just a basic understanding of statistical tables. Data swapping requires a good understanding of (demographic) microdata. Cell suppression requires knowledge of optimization techniques (e.g., linear programming). Some new sophisticated methods such as synthetic data require knowledge of ideas from Bayesian statistics and experience with statistical modeling. Even with some of the simpler methods, knowing exactly in what situations it is appropriate to use the method and how to fine-tune its use, requires experience. Similarly experience is needed to determine which method is best to apply in situations in which a number of methods are possible choices. Lately hybrid methods have become popular. Trying to create a coherent overview of these methods is a useful project for an agency that often needs to extend them to different situations. A course in disclosure avoidance would be a nice side benefit of such an effort. Such a course could be taught to new researchers in this field as well as those subject matter researchers who are often involved in disclosure avoidance issues.

Table of Contents

1.0 General Ideas about Disclosure Avoidance

The primary mission of statistical agencies is to collect data from single entities from some group (e.g., set of people selected from a survey frame or from a set of administrative records), aggregate it and then release it in the form of data for a wide range of subgroups. A subgroup might be as small as a few entities or as large as the entire group. The groups are generally those for which there is broad or strong public interest in having knowledge in the form of

statistical data. Some of the biggest such groups, are groups of all people, households, establishments, or companies for the entire United States. If these data are released for very narrowly defined subgroups of the full group, e.g., households on a block with few homes, there is the risk of revealing information about some (or all) of the individual entities that constitute the small subgroups. Such revelations are considered to be disclosures when the agency suspects knows or surmises that the revelations exceed what a data user can easily learn using free data sources.

Protecting all data releases from disclosures is a challenging goal and in order to implement it, we need to state the various forms in which statistical data will be released by an agency. The most common forms are microdata and tables. In this paper, we will focus on these two forms, while saying just a bit about other forms, such as statistical models.

In this paper, we will discuss recent developments in the protection of microdata and new ways to evaluate the risk of special tables (which are tables other than those planned for release). Then we will discuss an exciting theoretical development: the use of ideas from information theory to measure disclosure risk. This development has been led by computer scientists at corporate research institutes and at universities. The challenge will be to see how applicable it is to assessing disclosure risk of agency data products.

2.0 Recent developments with microdata: an introduction

At the Census Bureau, tables have been the dominant form of released data for the Decennial Census, the American Community Survey (ACS), and for almost all economic surveys and censuses. However, for ACS, public use microdata files (PUFs) are also released. In fact, for most demographic surveys conducted by the Census Bureau microdata is the only data product. When Census releases a large number of tables, involving a large number of combinations of the underlying variables, either as standard tables or as special tables (e.g., those requested by a user), tabular data may meet the needs of almost all users. However, it seems that researchers are increasingly interested in gaining access to the microdata underlying the tables.

Agencies are reluctant to release microdata, especially with the variables having a high level of detail. There are some traditional methods that coarsen the data, such as compressing many values of an ordinal variable into just a few categories, or numerical variables into just a few intervals. These coarsening methods increase the protection of the microdata but lower the data quality, and with it, the usefulness of the data to researchers. A compromise that is sometimes applied at the Census Bureau, is to retain much of the precision of the variables but release a small random sample of the microdata. The released sample typically consists of only a small percentage of the full sample of collected data.

The level of geography used in a PUF plays a key role in the protection of the data. In most PUF's, only very rough geographical information is supplied for each record. For Census Bureau surveys, there are pre-defined regions called PUMAs there are given for each record. The PUMAs for a given state form a partition of the state and are defined using recent decennial data; they are then fixed for decade until the next decennial. They are required to have at least 100,000 people. In some cases, even a super-PUMA (viz., a region with at least 400,000 people) is used as the geographical value. Use of PUMA's ensures that the geography variable, will not, by itself, provide much detail to a possible data intruder. However, if there are a number of non-geography variables that **are** released at a fine level of detail, the given record may be unique even in a large region, such as a PUMA or super-PUMA. Even if a data intruder does not know if a given record is unique in the population, he might assume that it is and try to link it to a record in another database. One way of making such linkages more difficult for the data intruder, is for the agency to release only a small percentage of the full collected sample; i.e., release only a random sample of the records formed from data collected from the survey sample.

If the agency were to release records with a geographic value that corresponds to a small (e.g., 10,000 people) or moderate (e.g, 50,000 people) region, an intruder would be even more likely to be able to perform linkages that could lead to disclosures. Here we assume the data intruder has access to a database that includes the same publicly known variables found in the Census PUF and this later database contains records for a large percentage of the population. If and when a match is made, the Census record is likely to disclose a number of additional data values, viz., those associated with variables in the Census record but **not** publicly known. . For example, there was study that showed that about 63% of the pairs (date of birth, sex) within a given zipcode are unique [ref: Golle]. This is an average over all zipcodes. So if these two variables appear in two microdata files for a given zipcode, one of which contains nearly the whole population of the zipcode, linkage of records leading to an expanded record for many persons is likely.

The increased interest in microdata has generated a lot of interesting disclosure avoidance research. For example, synthetic data, a process that involves using collected survey and generating a model from it, has captured the interest of academic statisticians [e.g., Reiter], as well as agency statisticians, [e.g.,Drechsler]. In addition, new ways of estimating the parameters of models (that describe the microdata), are being explored and may be almost as valuable to researchers as the microdata itself [Martin et al].

2.1 Comparing methods for creation of public microdata

Let us discuss some of the tradeoffs an agency needs to consider when deciding what protection method to use to protect publicly released microdata. The agency should consider older methods such as data-coarsening (e.g., topcoding) as well as new methods such as synthetic data. It is often the case that the data analytic effect of older methods is easily understood by researchers. For example, with data-coarsening, the researcher can often determine if the degree of coarsening will affect the type of analysis he is planning. As a bonus, the older methods typically are not computationally intensive. However, they often produce a microdata file that is deficient in meeting high-accuracy analytical needs of (primarily) researchers. There are two major computational steps involved in creating protected microdata usable by the public. One is protection of the data at some stage; the other is creating the public use file (PUF). For synthetic data those steps are combined into one. The main advantage of synthetic data is that when a researcher is using synthetic data, he is using data that has the same level of detail for all variables, including geography, as the original (edited) data.

One drawback of synthetic data is simply that it is synthetic. Specifically, it is creating microdata that are realistic, but not real. It is realistic in the sense that the data are consistent with the data that were actually collected, at least consistent with regard to those aspects of the data that were built into the data model used during the synthesis. But the microdata are not real since the records created do not correspond to actual survey respondents. Another drawback is the long time and effort it takes to develop a good data model.

2.2 A new method for creating protected microdata using multiplicative noise

There are other method being researched for creating protected microdata. One goal of this research is to find methods that are simpler to implement than synthetic data. Another goal is to provide a potential user of the microdata an opportunity to specify the main model that is used to create realistic data. One of these involves perturbing the microdata using multiplicative noise. See reference [Klein] for details. Here we present just a couple of the steps of this new method. A researcher would have to obtain the agency's EM (expectation maximization) software that requires the user to select a particular distribution to be used to model the microdata variable of interest. Currently, this involves merely selecting one of a small set of parametric models that the software is designed to handle. He then runs the software, with the specified model, against noise protected microdata released by the agency.

There is a potential stumbling block with this approach. An agency may be feel that such noisy microdata does not fully protect the data and thus will not allow it to be released publicly. So, perhaps the noisy microdata will be used along with the 'true' microdata as inputs to the EM software and the estimation of the model parameters will be done within the agency. The net result might be simply a good model for the variables involved or, possibly lead to the goal of 'model generated microdata' releasable to the public. Such modeling, in which the researcher can contribute to the modeling process, may produce microdata that is more useful to the researcher than a method in which models are chosen solely by the agency.

3.0 Recent developments with special tables

3.1 Introduction

Suppose a data user with interest in doing in-depth analysis of survey data (e.g., ACS data) finds that neither the agency's standard tables nor the public microdata are adequate for the analysis the user wants to perform. In this situation, the data user may request from the agency a set of 'special' tables that are close to what the data user needs for his research. Frequently, the agency will accommodate the request of the data user. Of course, this is possible only if the requested tables are capable of being generated from the survey microdata. (The agency may charge the user for the extra work involved in preparing the tables.)

However, on rare occasions, the agency may deny the request for special tables because of confidentiality concerns. What conditions could lead to such a denial ? This could occur if the set of requested tables are closely linked, and have some very low counts (say, 0, 1, or 2) in certain marginal positions. In this case, a data intruder might be able to generate a partial microdata record for an entity in the sample frame. When it is possible, this formation of a partial record can be done quickly and with little computational effort. It will not lead to a disclosure with direct identifiers, but it could easily lead to attribute disclosures. For example, it may lead to a fact such as: there is a person of sex S, of race R, and of age group A, residing on a given block who has income in the range R. If, from decennial census tables, it is known that there is a single person of type (S,R,A) on a given block or larger area, and if a neighbor knows of such a person, then the income of that person is revealed. Such attribute disclosures are likely when there are several 'known to neighbors' variables such as 'sex', 'race', 'age group', and 'means of transportation to work'. Sometimes this type of variable is also known to intruders who are not neighbors from publicly accessible databases.

Thus there is a dilemma. The agency will not release the requested tables if they are generated in the usual way; i.e. using the un-modified microdata. However, recent research has demonstrated that it is possible to modify the microdata, using some ideas from synthetic data analysis, to accomplish the two key goals (1) protect the data and (2) provide tables with values that are close enough to the true ones to meet the needs of the user. [Westat]

There are even some situations in which cell suppression can be used to protect a set of highly linked demographic count tables. There may be an overall significant loss of information from the data, but the loss in key tabular values (say, certain marginals) may be acceptable to the main users of the tables. [Massell, Hillmer]

There is a need for research on protecting special tables. This research needs to address two issues. (1) determining how an agency can quickly identify sets of linked tables that are likely to lead to (significant) attribute disclosures (2) determining the DA methods that are likely to produce a set of tables that are protected and have a type and level of information loss that is acceptable to the (requesting) user.

The identification issue (i.e. question 1) may be possible to automate; e.g., by writing software that requires only a list of variables to test whether generation of a risky partial record could easily result. However, the DA methods issue (i.e., question 2) may be much more challenging, at least that was the case with a particular set of ACS tables that focused on transportation questions [CTTP].

3.2 An agency online system to generate special tables and models

Online statistical databases are being used or at least considered as a way to let users quickly generate special tables that are restricted in certain ways but unrestricted in other ways. For example, the user must select from a set of pre-defined categorical variables. Thus having an 'age' variable, that creates a category for each year, may not be allowed. In such cases, the user will not have the same precision that access to microdata affords. Similarly, the user may be allowed to define the universe of the tables in a large number of ways, but there will be limitations. Also the tables are likely to be restricted to a small dimension k; this means that at most k variables can be crossed in any single table. Finally, the software may allow formation of a table but may not release it to the user, if the number of cells with a value of '1' exceeds an agency-determined threshold.

The Census Bureau's Microdata Analysis System (MAS) is undergoing development. There are some research papers that describe it and some of the statistical and computer science challenges it posed. [Lucero] [Zayatz] [Freiman]

4.0 Applying Information Theoretic Ideas to Disclosure Avoidance

In the mid 2000's, computer scientists developed some ideas that would allow an agency to be confident that individual data are not inadvertently revealed when the agency releases responses to statistical queries to a statistical database. In the field of statistical databases, this is called the problem of protecting the privacy of the database participants when performing statistical queries; ref [Micro]. Their work led to a new, (possibly) more precise, way to measure disclosure risk which they called 'differential privacy'. It is often challenging to apply this general definition to a particular data scenario. One important application to which it **was** applied is a geographic online database used for studying commuting patterns, called 'On the Map' [MAP], a joint development of Census and Cornell University researchers. For a paper that discusses why differential privacy protection mechanisms are not yet suitable for replacing those based on methods traditionally used with health care data (e.g., 'k-anonymity'), see [Dankar and El Eman].

In recent years, computer scientists, some at research divisions of major high tech companies; others at universities, have worked with statisticians to fine tune the notion of differential privacy so that it can incorporate data features commonly found in survey data, such as correlation among variables. Notions such as 'information leakage' and 'privacy leakage', some of which were developed prior to differential privacy, are now being used as a way to measure the gradual loss of privacy as responses are given to statistical queries or more directly from successive scheduled data releases. These notions are described in a very readable paper by Klarreich. See also the paper [Machana] in which the author gives a nice overview of the variety of methods for protecting microdata.

An excerpt from this latter paper is worth presenting:

"…there are two important considerations when designing a privacy protection method. First, the method must result in outputs that retains useful information about the input. Note that every privacy protection must result in some loss in utility (after all we are trying to hide individual specific properties). Hence, it is usually a good idea to list the types of statistical analyses that must be run on the data, and then tuning/optimizing the output to best answer those analyses. Some techniques assume an interactive setting, where a data analyst queries the datasets, and perturbed results are returned for these queries. Second, a privacy protection method should be simulatable -- an attacker must be assumed to know the privacy protection method. For instance, a method that reports the age of an individual (x) as [x-10, x+10] is not simulatable, since an attacker who knows this algorithm can deduce the age of

the individual to be x."

5.0 Conclusions

5.1 Determining the best available methods for a given scenario

In recent years, a wide variety of protection methods have been developed to protect microdata, and a lesser number developed specifically to protect tables. The overview books listed in the references (Hundepool et al, Duncan et al) do a good job of describing several methods for each major type of data product. When an agency is deciding which method to use to protect a given data product, it would be helpful if the decision could be made easier by merely listing certain features and parameters of the data product and then referring to some decision table that clearly delineates the tradeoffs involved among the suitable methods. However, confidentiality problems are so varied that the creation of such a decision table may not be a realistic goal. There may be no shortcut to the tried and true method of determining the 'best' protection method by simply working out the details for some test cases or for the specific dataset at hand. This often involves a lot of method specific software writing and a lot of computing. Namely, the agency will need to implement each method under consideration and apply it to each of a few realistic datasets, and then assess the analytic quality of the resulting protected data. It will also be necessary for the agency to develop an operational definition of disclosure risk for the data product in mind.

5.2 Attempt to add structure and theory to disclosure avoidance work

Disclosure Avoidance (or Statistical Disclosure Control) is a rapidly growing field in the sense that new methods are being developed each year. The effort to organize these disparate methods into a well-structured subject, as has been done with various areas of applied statistics or applied mathematics is quite challenging. However, we believe that it is worth the effort and could pay big dividends. At the very least, deeper understanding of the existing methods will result from trying to discover the common threads among the methods. Likewise it seems that gaining an understanding of the information theoretic approach to risk assessment would be useful for those people in agencies for whom disclosure avoidance is a major work component. We suspect that within a few years, enough experience will be gained using theoretical machinery to allow it to be applied to a number of standard agency data protection scenarios. If and when that is the case, studying the results of these analyses will help an agency decide the most sensible way to protect a given data product.

6.0 References

Anco Hundepool, et al. Statistical Disclosure Control, publ. by Wiley in 2012. In the Wiley Series in Survey Methodology. (the authors are 7 high level disclosure avoidance researchers from European statistical agencies).

George T. Duncan, Mark Elliot, Juan-Jose Salazar-Gonzalez , Statistical Confidentiality: Principles and Practice, publ. by Springer, 2011

Jorg Drechsler, Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation, publ. by Springer, 2011, Lecture Notes in Statistics, #201

Jerome P. Reiter, 'Data Confidentiality', http://www.stat.duke.edu/~jerry/Papers/wired11.pdf

Klein, Martin, 'Noise Multiplication and Multiple Imputation as Alternatives to Top Coding for Statistical Disclosure Control: An Overview and Comparison' JSM2013)

Lucero, Jason; L. Zayatz, M. Freiman, 'The Microdata Analysis System' (reports in 2009-2013)

Ciriani, V. et al, ' Microdata Protection', in Advances in Information Security, 2007, Springer ,http://spdp.di.unimi.it/papers/microdata.pdf

Massell, Paul B., Freiman, M. H., and Zayatz, L., "Data masking for disclosure avoidance", in 'Methods and Applications of Statistics in the Social and Behavioral Sciences', ed. N. Balakrishnan, John Wiley & Sons, ISBN: 978-0-470-40507-9, October 2012 Chapter 7:
https://collab.ecm.census.gov/div/cdar/Documents/Publications/Massell.Freiman.Zayatz.Handbook.article.pdf
The above article also appears in Wiley's online Encyclopedia of Statistical Sciences:
http://onlinelibrary.wiley.com/doi/10.1002/0471667196.ess1020.pub3/otherversions

[Massell; Hillmer 3] How cell suppression was used to protect the EEO tables.

Klarreich, Erica, 'Privacy by the Numbers: A New Approach to Safeguarding Data' ; Quanta (online) magazine, 2012, publ. by Simons Foundation. https://www.simonsfoundation.org/quanta/20121210-privacy-by-the-numbers-a-new-approach-to-safeguarding-data/

[Micro] Microsoft Research, 'Database Privacy' (long list of articles by Cynthia Dwork and Frank McSherry and others). 2011 http://research.microsoft.com/en-us/projects/databaseprivacy/

Dwork, Cynthia, and Adam Smith, 'Differential Privacy for Statistics: What we Know and What we Want to Learn' Journal of Privacy and Confidentiality (2009) 1, #2, pp. 135-154
http://repository.cmu.edu/cgi/viewcontent.cgi?article=1008&context=jpc

[MAP] Machanavajjhala, Ashwin 'Privacy: Theory meets Practice on the Map'; conf.
http://www.cse.psu.edu/~dkifer/papers/PrivacyOnTheMap.pdf

[MACHANA] Machanavajjhala, Ashwin , 'Big Privacy Protecting Confidentiality in Big Data'
http://www.cs.duke.edu/~ashwin/pubs/BigPrivacyACMXRDS_final.pdf (in ACM digital library)
http://www.truststc.org/pubs/463/icde2008-privacy.pdf

Danker, Fida K., and Khaled El Eman, 'Practicing Differential Privacy in Health Care: a Review', Transactions on Data Privacy 5 (2013) 35-67. http://www.tdp.cat/issues11/tdp.a129a13.pdf

Alvim, Mario, et al; 'Differential Privacy: a Study of Utility and Min-Entropy Leakage'
http://homepages.laas.fr/mkilliji/APVP2011/Site/Programme_files/Article_12.pdf

[CTTP] {some paper from CDAR or WESTAT describing the CTTP problem & soln}

[Westat] { Westat reports: how microdata was modified; then used to generate CTTP (Transp. Planning) tables from ACS data}… Tom Krenzke, Jane Li, et al.

Golle, Phillippe, 'Revisiting the Uniqueness of Simple Demographics in the US Population', http://www.truststc.org/wise/articles2009/articleM3.pdf