

# Assessing the Statistical Disclosure Risk of a Demographic Microdata File

Paul B. Massell

U.S. Census Bureau

(e-mail: paul.b.massell@cmail.census.gov)

## Introduction

There are two recent developments related to survey data dissemination that may be increasing the risk of disclosure of respondent data. One is that statistical agencies are now releasing more microdata files than previously, partly in response to the urging of researchers needing the data for precise analytic work. For example, some data rich files with possibly high disclosure risk, that have been at least considered for public release, contain longitudinal survey data that has been linked to administrative data. The other is the development of numerous databases containing personal information of use to businesses for credit assessment and targeting of advertisements. Some of these databases have records that can be matched to survey records since the demographic data items collected are often similar for databases and surveys. Since the databases in general have explicit identifiers in their records, anyone with access to such a database has one important tool needed for data snooping. Furthermore, some of these databases with matching potential are being placed on the Web, which of course greatly increases the ease with which a data intruder can access the data. In this paper we have two goals. One is an attempt to formalize the process of assessing disclosure risk a bit beyond what is done in two recent excellent surveys of the field (ref: WdW, EURO). Another is to report how we applied some of these general ideas to the assessment of disclosure risk to a particular microdata file along with suggestions for reducing the risk. The particular file comes from the American Housing Survey but most of the analysis of those data can be applied to other survey microdata.

## 1. The Microdata File and its associated Population File

Consider a demographic microdata file  $M$  to be a set of records, one for each respondent. A respondent may supply data for only himself or for some or all members of the household. The records in the file describe either a person or a household. In either case there may be missing data items (variables). The respondents form only a subset of the set of persons or households sampled. The analysis of non-responses, either entire records or item non-response is part of response error analysis. The set of sampled persons or households is, for a non-census survey, a proper subset of the entire sampling frame for the study. Let us consider a theoretical construct, the set of records with no missing items for all persons or households in the sampling frame. Let us denote this construct by  $\text{Pop}(M)$ , and call it the population file associated with the file  $M$ .

An important issue, that we will ignore initially, is the analysis of measurement error. For simplicity, we will assume for now that all the non-missing values in the records of both  $M$  and  $\text{Pop}(M)$  represent the true values of the variables. Disclosure analysis differs from other statistical topics in

the role that measurement error plays; specifically response variation is often more important than response bias. This is because matching is at the heart of many disclosure techniques and if a variable is consistently reported with the same bias on both the survey file and potential matching files, the bias probability will have no effect on the ability of an intruder to match records. However whenever a variable ( e.g., income) is reported differently on different questionnaires, this variation will, in general, make matching more difficult. In addition to response variation there are other sources of variation that the disclosure risk analysis may need to consider in order not to overestimate the probability of a data intruder's ability to match records. However, even with these obstacles, a data intruder may be able to match records if he has at least roughly accurate knowledge of a large number of variables (ref: W).

## 2. The notion of a key

A key  $K$  for a microdata file  $M$  is a subset of the variables defined for the records in  $M$ . In disclosure analysis, a key is often chosen to contain variables that are accessible from many different files; e.g., keys often include basic demographic variables. Thus a typical situation is one in which the key is known for  $M, M_2, M_3, \dots$  where each microdata file is generated by a survey and none of the files contains  $\text{Pop}(M)$ . A less common situation, but nevertheless important, is one in which  $M$  represents a non-census survey, so  $M \subset \text{Pop}(M)$  but there may be a 2<sup>nd</sup> file  $M_2 \supset \text{Pop}(M)$  on which  $K$  is known. Thus, even though  $M$  represents only a subset of  $\text{Pop}(M)$ , the key is known for all of  $\text{Pop}(M)$ .

Example: Suppose  $K = (\text{sex}, \text{race}, \text{hispanicity}, \text{age})$ . Then this key is known on the Census persons file  $M_2$ . Since for any persons survey file  $M, M_2 \supset \text{Pop}(M)$ ,  $K$  is known for  $\text{Pop}(M)$  but most of the non-key variables are not known for  $\text{Pop}(M)$ .

Since there are a large number of files that could be used for matching records of  $M$ , it is important to carefully select the keys to be investigated. There are a few general statements one can make about which keys are most important for disclosure risk analysis. The basic demographic variables such as sex, race, age, are likely to be on many matching files. Salary or income may be on some such files. At the household level, number of persons in household, and number of children, and number of persons who have drivers' licenses are often present on external files. Since much data are collected with the goal of being sold to advertisers, variables that reflect the financial status of individuals are likely to be known for most people in the population. These include salary, value of house or rent for apartment, and use of credit for purchases. It should be mentioned that the intruder can use different combinations of keys to re-identify different subsets of records.

## 3. Types of disclosure

### A. *Re-identification Disclosure*

This type of disclosure is what is usually meant by the term 'disclosure'.

This type requires the ability to assign an identifier, e.g., name, address, telephone number, or Social Security Number (SSN), to one of the records in  $M$ . We say each such record has been 'disclosed'. Since file  $M$  does not contain such identifiers, one must obtain this information from another source. The other sources could be

- (1) a survey file with identifiers
- (2) a database file with identifiers
- (3) a partial list of respondents to M's survey
- (4) knowledge of neighbors.

In cases (1) and (2) the risk of disclosure is open to a wide class of intruders if the files are publicly accessible and inexpensive. However, even if the file is proprietary and only accessible to a few employees, of say, a credit reporting agency, we still need to consider this data as posing a disclosure risk, albeit to a limited number of data intruders. In case (3), such a partial list is sometimes obtained by word of mouth and may consist of identifiers of respondents and a few key (or other) variables in M that allow quick identification of the record of each person on the list. Thus, a disclosure of all previously unknown variables in M results for all listed persons.

Database files with identifiers now appear on the Web. Examples of these are driver license files for many states (ref: NIC ). Newspaper articles have recently appeared that describe data brokers who match records from a variety of online and proprietary sources to create extensive records on individuals (ref:WP1, WP2).

For case (4), suppose the key consists of easily acquired variables. These are variables such as age, race, sex, and hispanicity which are often, at least approximately known, by neighbors of any given resident. Suppose we know there is a record with key = i that is unique in some geographically defined (population) sub-file, e.g. in a census file for a particular city. Suppose there is a survey record for this unique sub-file record. If the sub-file's region is small, and the key is easily acquired, then the person with the unique record may be known to his neighbors. Then all the data for this resident's survey record can be obtained by any neighbor who knows that there is a survey record for this resident.

#### *B. Attribute (or prediction) disclosure*

The term "attribute disclosure" is used by several authors but there is some variation in its usage (ref: LA). Our usage follows that in WdW, p.15. "Prediction disclosure (or attribute disclosure) occurs if the data enable the attacker to predict the value of a sensitive variable for some target individual with some degree of confidence. For prediction disclosure it is not necessary that re-identification has taken place." We can extend this definition to include disclosure of a sensitive variable for a group of (one or more) people usually identified by basic demographic variables such as sex, race, age, and hispanicity. This is a more subtle type of disclosure than re-identification. Whereas re-identification disclosure has a binary outcome, viz. disclosure of a respondent's entire record or not; the outcomes of attribute disclosures are best described in terms of the **degree of disclosure**. Attribute disclosure occurs if one can determine from M alone that a group has a distribution on a sensitive variable that is significantly different from the population's distribution and the group's distribution is revealing in that it shows a limited range for a continuous variable or a limited number of categories for a categorical variable. When we say 'determined from M alone' we are including **estimation** of the group's population distribution from the sample that appears in M.

#### **4. Measuring disclosure risk**

### *A. The notion of a disclosure risk measure DRM*

So far we have described types of disclosure but have not discussed how to measure the vulnerability of a microdata file  $M$  to such disclosures. Before discussing the particular DRM with which we have done all of our computations, let's mention a general problem in applying any DRM. Suppose  $M$  is a microdata file containing the survey sample for a national survey. If the sampling fraction is small, say  $1/1000$ , there is probably little reason to worry about disclosures at the national level. The greater concern is the release of subfiles of  $M$  that represent small regions of size, say, 100,000 or 200,000. In the process of estimating risk for the regions, it is helpful to assume that the file  $M$  may be treated as a sample file for the regions of interest. When this assumption is justified, it allows one to use "extension methods" on  $M$  that are valid for regional populations that are say, a small multiple of the size of  $M$ . For example, if  $M$  has 50,000 national records, we may assume for the purpose of DRM estimation that those records may be viewed as sampled from a region  $R$  of population  $\text{Pop}(R)$  of size 200,000. Even using that bold assumption, extending DRM estimates from  $M$  to  $\text{Pop}(R)$  can be challenging.

### *B. An important DRM based on uniques*

Let  $M(R)$  denote the set of microdata records for a region  $R$  formed either by selection from the national file using a geographical identifier or from a separate sampling of  $R$ . Let  $\text{Pop}(R)$  be the theoretical construct for the region  $R$  analogous to  $\text{Pop}(M)$  described above. One reasonable DRM is the fraction of records in  $\text{Pop}(R)$  that are unique with respect to a given key. We call this DRM the "fraction uniques DRM." If the sampling fraction for region  $R$  is small enough, say  $1/1000$ , to assume independence of the "uniqueness property" among the records in  $M(R)$ , one could use this DRM to approximate the probability that at least one record in  $M(R)$  will be unique in  $\text{Pop}(R)$ ; viz.,  $1 - (1 - \text{DRM})^t$  where  $t = \text{size of } M(R)$ . The estimation of this DRM involves two steps. Assuming the use of the national file  $M$  can be justified by some homogeneity-type argument, and that the size of  $M$  is less than  $\text{Pop}(R)$ , we first calculate the number of uniques in  $M$ , say, using a simple SAS program. (If the size of  $\text{Pop}(R)$  is less than  $M$  then one begins by taking a sample of size  $\text{Pop}(R)$  from  $M$ ). The next step can often be challenging (as mentioned in section A above.) One needs to use some method of estimating the fraction of uniques in  $\text{Pop}(R)$  from knowledge of uniques in  $M$  and possibly other equivalence classes in  $M$ . The simplest such extension formula involves only the uniques in  $M$ . More accurate estimates however can be derived by using the number of  $k$ -classes for  $k=1,2,\dots,L$  where  $L$  is the size of the largest equivalence class that exists in  $M$ . See GZ and BKP for the development of more precise extension formulas.

We should mention that our use of a DRM that estimates uniques is currently lacking a method of constructing a confidence interval for the estimate of fraction uniques in  $\text{Pop}(M)$ . Specifically we can test the bias of the estimate by comparing it to those generated by more refined methods, but a measure of uncertainty needs to be developed.

### *C. Formalizing the idea of group-type attribute disclosure using information theoretic measures*

The ideas below are one way of formalizing the idea of group disclosure. We begin with a motivating question, a general discussion, and an example.

When we have disclosure about a group, the observer is learning things about the group. How does

"disclosure information" differ from other types of information?

Example: Suppose green males in a (small) regions have a markedly different income distribution from that of the total population. Suppose this is known from a census or is inferred from a sample. Now if the green males' income distribution is bounded, we say "all green males have income < 20 K in this region." This is both informative and a group disclosure. If the green males' incomes tend to be lower but occasionally are not, that situation may be describable by a probabilistic statement that is nearly as informative as the (deterministic) boundedness statement given above.

$$\text{Prob}(\text{green male income} > 20 \text{ K}) \ll \text{Prob}(\text{male income} > 20 \text{ K})$$

(where " $\ll$ " means "much less than").

The following definitions are part of our attempt to formalize the idea of the amount of additional information about a group that can be provided by the subset of M that corresponds to the group. They use the idea of a (statistical) divergence (function) which is a weak type of distance function which is not required to satisfy the triangle inequality that an ordinary distance function (i.e., a metric) must satisfy.

Definition: The "degree of divergence with respect to the variable X for the group defined by K = i" is the statistical divergence between P(X) and the conditional P(X|i). If we let Y = X|i, the statistical divergence between X and Y, denoted J(X,Y), measures the difference in the distributions of X for the (1) entire population and (2) the sub-population defined by K = i.

The above definition is abstract and does not specify the form of the divergence.

Assume X is a categorical variable; either initially or after discretization of a continuous variable. Let pf(X) denote the probability function of X. Let  $\text{pf}(X) = \{f_1, f_2, \dots, f_n\}$   
Let Y = X|i be X restricted to the group defined by key = i. Let  $\text{pf}(Y) = \{g_1, g_2, \dots, g_n\}$   
Then we have a definition of divergence, called the J-divergence, that is based on the idea of Shannon entropy from information theory (ref: A,K,ESS):

$$J(X, Y) = \sum_{j=1}^{j=n} (f_j - g_j) \cdot \log(f_j / g_j)$$

This is only one possible definition of divergence. See ESS for an overview of this topic.

Definition. Let X be any random variable defined for records in a microdata file. Let Y=X|i be the restriction of X to a subpopulation defined by the value i of a key. If the divergence J(X,Y) is large and if Y is concentrated on an informative subset of the range of X then the **file provides a disclosure about group i's values of X.**

If X is a categorical variable, an "informative subset" is one that has  $f_j > 0$  and  $g_j = 0$  for some interesting values (categories) of X, i.e. Y's range excludes an interesting value that is in X's range. If X is an ordinal variable, we require that  $f_j > 0$  and  $g_j = 0$  for either the smallest or largest value in the range of X i.e., Y has a smaller upper bound or a larger lower bound than does X. This definition provides a basis for a measure of disclosure risk for group-type attribute disclosure.

If the microdata file  $M$  represents a survey,  $X$  may be viewed as defined on  $\text{Pop}(M)$ . All of the above definitions extend to this case but now the  $f_j$  and  $g_j$  are no longer probability function values but simply observed relative frequencies. Having only a survey, we are faced with the usual problem of how to make inferences about  $X, Y$ , and  $J(X, Y)$  for  $\text{Pop}(M)$  based on frequencies on  $M$ . I believe that research needs to be done on the extension problem for attribute disclosure. However there are two special cases in which extension is not required.

(1) a census:  $M = \text{Pop}(M)$

(2) a survey for  $M$ , i.e.,  $M \subset \text{Pop}(M)$  and a separate file  $M_2$  that contains a census for the key.

That is,  $X$  is known only on  $M$ , but the key  $K$  is known on  $\text{Pop}(M)$ . Knowing  $K$  on  $\text{Pop}(M)$  allows us to know when all of the  $K = i$  records in  $\text{Pop}(M)$  are contained in  $M$ .

## 5. Comparing disclosure risk measures for national and regional files.

Let  $M$  be a microdata file for a national survey. Let  $R$  denote a regional microdata file where regional simply means sub-national.  $R$  may or may not be a subset of some national file.  $M$  must contain some geographical variables in order to select  $R$  from  $M$ . In that case, the sampling fraction (actually response fraction) for  $R$ , i.e.  $R$  in  $\text{Pop}(R)$ , is likely to be nearly equal to the sampling fraction of  $M$  in  $\text{Pop}(M)$  if  $R$  is sufficiently large. Often, however, the regional file is derived from a sample that is separate from the national survey; in this case the sampling fraction (and other survey features) may be different for the two surveys.

In the latter case, when assessing disclosure risk, one must compute the fraction uniques DRM on  $(M, \text{Pop}(M))$  for the national file and on  $(R, \text{Pop}(R))$  for each regional file  $R$ . Any file for which the corresponding DRM is high should not be publicly released.

When determining what regional file  $M(R)$  to form, the data provider should be aware of the following fact. The basic idea is that, **on average**, the fraction uniques DRM **decreases** as the size of  $R$  **increases**. This may be shown by means of a simple statistical argument. We emphasize, "on average" because it is easy to construct examples where the rule does not hold for a particular subset of a  $\text{Pop}(R)$ .

Applying this argument leads to the "rule of thumb": the smaller the subfile the higher the fraction of uniques. For this reason, it is common practice at some statistical agencies to impose a minimal regional size  $|\text{Pop}(R)|$  in order for the corresponding file  $R$  to be released. This minimal size is stated as a number of persons, e.g., "100,000 persons". If one wants to maintain a consistency between the minimum for the number of persons and the minimum of the number of households, one may require a minimum of say 40,000 households, since that corresponds roughly to 100,000 persons.

What's special about the geography variable ?

There are a number of reasons why regional files are produced. They meet the needs of public officials to have information about their jurisdictions. Similarly, business people may want to have information about the areas they "cover." Researchers often like to compare cities or regions. However, in relation to the disclosure problem, the existence of such files makes group disclosure

more likely. The risk for a re-identification disclosure is increased by the existence of a geography variable on possible external files that can be used for matching. With a regional survey file, and one or more matching files for the same region, the data intruder's work is easier than if only national files were available. The regional disclosure problem is more likely to lead to a disclosure for a given amount of effort by the intruder.

## **6. Ways to reduce the disclosure risk**

Let us list a number of measures commonly used to reduce the disclosure risk.(ref: EURO; WdW)

### *A. Increase the minimum population size from which a regional sample is drawn.*

A population size threshold commonly used is 100,000; this may be increased to, say, 200,000.

The discussion above explains why increasing the minimum size reduces the fraction of uniques measure for disclosure risk.

### *B. Top and/or bottom code some of the key variables.*

With ordinal variables the tails of a continuous distribution or the lowest and highest categories of a categorical variable, often have fewer cases than the mid-range of the distribution. By top and/or bottom coding, we are eliminating these low probability categories.

### *C. Coarsening of key variables; collapsing (or broadening) of categories(ref: M, WdW).*

The more general term 'global recoding' includes this method. By 'coarsening' we mean combining of categories to create fewer and higher probability categories. For ordinal variables, one combines adjacent categories. Top and bottom coding may be viewed special case of this method. There is obviously the potential for a significant decrease in data utility if the coarsening proceeds beyond a certain point.

The determination of an optimal tradeoff between information content and disclosure risk is explored in some papers, including (ref:M) in which Shannon entropy is used to measure information content of the file.

### *D. Local suppression of key variables (ref: WdW)*

Local suppression involves blanking out some extreme variable values that are likely to contribute significantly for the disclosure risk for the record involved. The author does not have experience with method but apparently it is commonly used in certain European statistical agencies (ref: WdW, EURO). Information loss due to local suppression could be explored using Shannon entropy.

### *E. Swapping*

Within a small geographical unit (but not the smallest), we switch the majority of variable values for two records that have certain key variable values in common. Since this switching is done within a geographical unit, all statistics at the level of that unit and higher are unaffected. However, by perturbing the true data values in this way makes matching either record to an external file more difficult. Specifically the geography information at the smallest geographical unit is altered for a swapped record. For example, if two Hispanic males of a certain age live in the same block group but not in the same block, swapping their records has no effect on statistics at the block group level

but probability will have a minor effect on statistics at the block level. The change of the block for the two records may make disclosure less likely. It also makes attribute disclosures involving groups of size one less harmful in that the variable values “revealed” are likely to be inaccurate. This latter case is not the ideal outcome of swapping, i.e., for a data intruder to feel he knows the income of some person from his file record when in fact that file record belongs to someone else. A better outcome of swapping, is to put uncertainty into the mind of a data intruder; viz. that he doesn’t know if a given record is accurate. Hopefully, this uncertainty will discourage him from attempting a disclosure in the first place.

*F. Perturbation: adding a small amount of a noise to variable values.*

Using a symmetric-about-zero random number distribution with truncated tails, one adds noise to all values of a given variable in a way that does not affect averages very much but may sufficiently perturb the data at the microdata level to make disclosure more difficult. Correlation between two variables, one or both of which have been perturbed, may likewise be perturbed, but probably only slightly. (ref: KIM, F)

## **7. Estimating the disclosure risk of a national American Housing Survey (AHS) file. (ref: M2)**

*A. The frame and type of variables.*

The frame consists of all housing units in the U.S. in a given year whether occupied or not.

The sampling fraction is about one in 2000 units, although certain regions have a higher rate. For each occupied unit basic demographic variables are collected on each person in the household and hundreds of details about the unit itself, viz. number of rooms of each type, details about the construction of each part of the unit, expenses for rent or mortgage and for additions to the unit, etc. The demographic variables collected are similar to those for the Census of Population and Housing and the Current Population Survey.

*B. External files that could be used for matching records.*

There is an interest in assessing the disclosure risk for the AHS surveys. The emphasis is currently on estimating risk for disclosure by re-identification. To do this we would like to know, ideally, what external files have at least a few variables in common with the survey file and are of potential use for re-identification. We have knowledge of the types of variables available on such files but have not, as yet, actually acquired any of these files. One source of this knowledge is literature distributed by companies describing their proprietary files, e.g. credit reporting agencies; another source is newspaper reports about personal data (or information) brokers.(ref: WP1) In some cases, files are available at little or no cost on the Web; e.g., drivers' license files (ref: NIC).

*C. Choosing keys and assessing the uniqueness measure*

To date, we have focused on the persons subfile of the national AHS microdata file. This is because we feel many of the variables found there are likely to be found on external files. We explored the key = (age, sex, salary (\$2000)) where salary (\$2000) denotes the salary recoded to \$2000 intervals above. This makes sense because salary is likely to be reported only approximately. The persons file had 102,761 records; it had 847 uniques; this is equivalent to 0.824 %. To this limited equivalence class data we apply an extension formula. We choose what is probably the simplest such formula.

This is basically the subsampling algorithm presented in ref: GZ, p.41. We have derived a slightly simpler formula by making an additional weak assumption. The formula is applied at the end of the following algorithm. Denote the given population P's size by  $N_p$  and the number of uniques by  $U_p$ ; thus  $N_p = 102,761$  and  $U_p = 847$ . Suppose we want to determine the value of this DRM for a population of size  $N_h$  where  $N_h > N_p$ . Then we choose a sample of size  $N_t$  where  $N_t/N_p = N_p/N_h$ . Let  $U_h$  denote the estimate of the number of uniques in a typical population of size  $N_h$ . Let  $f_1 = U_h/N_h$ ,  $f_2 = U_p/N_p$ , and  $f_3 = U_t/N_t$ . The subsampling method is based on the simple idea that the fraction of uniques satisfies  $f_1/f_2 = f_2/f_3$ . This implies  $U_h/N_h = (U_p/N_p)^2 / (U_t/N_t)$ . This extension formula allows us to estimate the fraction of uniques for any population file that, in theory, contains the given microdata file. The validity of the estimate was explored numerically in GZ by comparing it to a more precise estimate; for the example presented there, the subsampling formula yielded nearly equivalent values. For our work, we assessed only the variability of the estimate by choosing 10 subsamples of size  $N_t$ . The variation was not great and we felt confident in taking the average of the ten estimates of  $f_3$  for use in the formula. This method yields as estimate of 0.394 % for fraction uniques for a population of 200,000 and only 0.070 % for a population of 1,000,000. Ref GZ suggests that estimating fraction uniques for populations more than ten times the size of the given population is not reliable.

The percentage of uniques becomes much larger when one combines all the persons data for all persons living in a housing unit. In general, such a group constitutes a family, and there are likely to exist external databases that have the same variables for all adults in a family, or all holders of a driver license. In fact for the key  $K =$  (age and sex of each person at least 16 years old in a housing unit) about 20% of the housing units are unique. When one coarsens "age" into five year intervals, the percentage is still about 5%. No decision has been made regarding recoding of present variables, but it is likely that age and other variables will be coarsened in order to significantly lower the disclosure risk due to the presence of unique records.

## **8. Conclusions**

Our experience with the AHS data leads us to believe we should seriously consider releasing only modified data for basic demographic variables and other variables that could be used as key variables for matching to external files. Since these variables appear on many surveys, perhaps instead of performing a disclosure analysis for each such survey, a generalized analysis should be done in the near future which determines which demographic variables can be released and in what detail. There are likely to be objections from some researchers who will claim that data utility is being reduced. However, if we do not reduce the information content and disclosures do occur and are publicized, the political repercussions may lead to a mandating of a more severe reduction in information content for future releases, including at the extreme, not allowing public release of certain files. Concern with disclosure is already leading to an increase in the number of microdata files that are released only to researchers who are willing to sign a restricted access agreement. This is a legal document (a data licensing agreement) in which the researchers agree to handle the data safely, to allow any publications to be reviewed by the agency providing the data, and to restrict access to a small designated list of users at the licensed research facility.

## References:

- A: Ash, R., Information Theory, Wiley 1965.
- BKP: Bethlehem, JG, Keller, WJ, Pannekoek, J Disclosure Control of Microdata, JASA, March 1990, v.85, pp38-45.
- B: Biemer, Groves, Lyberg, Mathiowetz, Sudman, Measurement Errors in Surveys, Wiley, 1991
- C: Cox, L. Statistical Disclosure and Disclosure Limitation, JPSM course notes, Dec. 1998
- ESS: Encyclopedia of Statistical Sciences, Wiley 1982-1998.
- EURO: Manual on Disclosure Control Methods, EUROSTAT, Brussels, 1996.
- F: Fuller, W.A., "Masking Procedures for Microdata Disclosure Limitation," JOS,9 (1993)
- GZ: Greenberg, BV, and Zayatz, LV, Strategies for Measuring Risk in Public Use Microdata Files Statistica Neerlandica , 1992, vol 46, nr. 1, pp. 33-48
- KIM: Kim, J.J. "A Method for Limiting Disclosure in Microdata based on Random Noise and Transformation," ASA Proceedings of Section of Survey Research Methods, 1986, 303-308.
- K: Kullback, S. Information Theory and Statistics, Dover, 1997.
- LA: Lambert, D., "Measures of Disclosure Risk and Harm," JOS,9(1993) 313-331
- LZ : Linhart, H., Zucchini, W., Model Selection, Wiley, 1986
- M: Massell, P B, Deterministic Modifications of Microdata: Balancing Disclosure Risk vs. Information Loss (to be published in Proceedings of the section on Survey Research Methods, American Statistical Association, 1999)
- M2: Massell, P B, Disclosure for Microdata Files: Assessing and Reducing Risk (with special application to the American Housing Survey), Informal Paper, 1999
- NIC: <http://www.technosearch.com/naked/directory.htm>
- W: Winkler, W.E., " Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata", Research in Official Statistics, 1, 87-104. 1998.
- WdW: Willenborg, L, de Waal, T., Statistical Disclosure Control in Practice, Springer, 1996
- WP1: Washington Post 3-part Series on Data Brokers, March 8-10,1998
- WP2: Washington Post article on citizen data collected by the federal government, June 27, 1999

## Appendix: SAS programs used in this work.

- \* This code implements the subsampling algorithm described above
- \* Selecting a subsample of size that allows estimation of fraction uniques for a pop. of 200,000
- \* Note  $(200,000 / 102,761) = (102,761 / 52,799)$
- \* Finding the number of uniques for a key (sex, age, recoded salary)
- \* Forms a subfile with these unique records
- \* Macro performs these steps for a designated number of times ("howmany"); this is desirable so one \* can use an average value of the fraction of uniques in a subsample of given size

```
%macro resamp(howmany);  
%do jmac = 1 %to &howmany;
```

```
data samp_a (drop = n k);
  set personc ;
  retain n 102761 k 52799 s 0;
  samp = 0;
  if(s lt k) then do;
    rate = (k-s) / (n - (_N_ - 1));
    x = ranuni(0);
    if (x lt rate) then do;
      samp = 1;
      s = s + 1;
    end;
  end;
run;
```

```
data samp_a ;
  set samp_a ;
  if (samp eq 1) ;
run;
```

```
proc freq data = samp_a;
  table age * sex * salrec / noprint missing out = subflat;
run;
```

```
data subuniqs;
set subflat;
  if count ne 1 then delete;
run;
%end;
%mend resamp;
```