Disclosure Risk Assessment for Survey Microdata

Trivellore Raghunathan and John Van Hoewyk

Institute for Social Research

University of Michigan

Ann Arbor, MI 48106-1248

Abstract

Limiting disclosure of a respondent's identity, and hence his/her responses to questions asked in a survey, is an important issue. Though a variety of statistical models have been used to assess the risk of disclosure, limited investigations have been reported in the literature assessing the risk in practical contexts. This paper reports on the results of two experiments attempting to glean the identity of respondents from several national surveys using a commercial database with names, addresses and several key variables common between the survey and commercial data sets. Furthermore, the measurement error properties in the common variables were also evaluated by comparing the survey reports with those in the commercial databases. This research suggests that concerns about risk of disclosure may be overstated and perhaps, concerns should be channeled towards protecting the confidentiality of responses to those situations where the intruder knows that the subject participated in the survey. Even here, the measurement error properties of the responses from the respondent and assumed knowledge of the intruder may buffer any concerns about risk of disclosure.

Key words: Confidentiality, Disclosure Limitation, Survey Research, Privacy and Protection, Public-Use Data

1. Introduction

Demands for micro level data for statistical analysis purposes, especially if collected using taxpayer funds, have been growing steadily. Such use of data can be quite beneficial for a society to render informed public social, economic and health policies. However, such free access to data raises concerns about fulfilling the pledge of confidentiality promised by the agency when collecting the data from the respondents. The conventional wisdom indicates that when more variables are released for statistical analysis purposes the risk of disclosing the respondent's identity, accidental or otherwise will also increase.

In response to such a threat, several approaches have been developed for statistically masking the data or setting up enclaves with strict rules of access. In the former approach, it is possible that the statistical properties of the data may be altered and thus introducing bias in the survey estimates (Liu 2003). The latter approach essentially limits access to a select few individuals and goes against the notion that broad dissemination of survey microdata is essential for an open society.

The risk of disclosure assessment has two parties, a data agency that collects and disseminates the data for statistical analysis purposes and an intruder who misuses the data to obtain the identity of respondents. There are two types of intruders (Duncan and

Lambert (1989). The first type of intruder (Type I) has a target person in mind and *knows* that he/she participated in the survey. An example of this scenario is a parent-intruder who knows that his daughter participated in a sexual behavior survey and is trying to find out about her reported number of sexual partners. The intruder knows certain attributes of the target which are also part of the data set released by the data collecting agency. The potential intruder sifts through the data by matching the attributes. In this sifting process, if the intruder finds a unique match then the target is potentially identified. If the intruder finds many matches during the sifting process then the target is protected unless there is some other information in the survey (e.g. date of the survey) that could be used to reduce the set to a unique respondents in the survey. Thus, in this case any sample unique based on a set of attributes that will be known to an intruder may have unacceptable risk of disclosure.

The second type of intruder (Type II) does not have a target in mind but has access to a large population database with identifying information and some common variables, *X*, available in both a population database and the survey data set. An intruder then matches the two databases on X. Now suppose that n_x and N_x are the number of records in the survey and population databases for a particular value $X = x$. The risk of disclosure in this case is governed by the joint distribution of (X, n_X, N_X) or, specifically the conditional distribution of (n_X, N_X) given *X*. For example, if $n_x = 1$ and $N_x = 1$ (that is, we have a sample unique and a population unique), and assuming that there is no measurement error in *X* (that is, the value of *X* recorded in both databases for the same individual are identical) and no coverage error (that is, the respondent under question is

in the population database with probability 1), then the corresponding respondent is identified. For a large N_x, the risk of disclosure is minimal even if the respondent is sample unique. In general, the risk of disclosure may be quantified in probabilistic terms using the ratio $p_x = n_X / N_X$.

Risk assessment in a practical context is a complex process. There are several key factors that affect the disclosure risk in both types of intruder problems. First, is there measurement error in *X*, wherein *X* available in population database will differ from the survey data on the same individual? Measurement error in the attributes may lower the risk of disclosure. For example, if the attributes of the respondent known to the intruder does not match with the attributes given by the respondent to the survey then the intruder obtains a false match. However, as far as the disclosure risk assessment is concerned the respondent has not been identified.

The second factor affecting the risk assessment for Type II intruder is the coverage error in the population database where not all subjects in the sampling frame used in the survey are included in the population database. The third key factor is the nature of *X* itself. The risk assessment evaluated, for example, in Sweeney (1998, 2000) assumes geographical details such as zip code, the exact data of birth and some other information that are not usually released as a part of the survey data. All these factors may lower the risk of disclosure in meaningful ways. On the other hand, a respondent to a survey and an intruder may collude to tease out subjects among the potential matches, thereby altering both n_X and N_X which may actually increase the risk of disclosure.

Statistical models to estimate the risk of disclosure for a given data set for a Type II intruder problem have been in vogue for some 20 years. Duncan and Lambert (1989) is perhaps pioneering work in this area. Additional methods for estimating the risk of disclosure includes Bethlehem et al (1990), Lambert (1993), Fienberg and Makov (1998), Skinner and Holmes (1998), Skinner and Elliot (2002), Elamir and Skinner (2006), and Reiter (2005) among others. Fuller (1993) discusses estimation of disclosure risk in the presence of specific type of measurement error when *X* has a multivariate normal distribution. However, in most circumstances *X* will be treated as categorical variables for matching purposes. Most of these are statistical methods for estimating the risk using the current survey data with certain assumptions about the population characteristics. This paper reports on two identification experiments that assess the risk of disclosure in a practical setting as well as assess the measurement error properties of variables used as matching variables and their impact on disclosure risk.

2. Setup

Suppose that the data collection agency collects $\{I_i, x_i, y_i; i = 1, 2, ..., n\}$ on n subjects where, for subject i, I_i is the identifying information (such as names, addresses, telephone numbers, Social security or driver license numbers), x_i is a vector of key demographic variables and y_i is a vector of survey variables of interest. The de-identified data is $\{x_i, y_i\}$ and is released as public-use microdata. Suppose that an intruder has an administrative data set $\{I_j^*, x_j^*; j = 1, 2, ..., N\}$ on N subjects with a collection of identifying

information I_j^* and key demographic variables x_j^* on subject *j*. (The superscript * is used to differentiate administrative data from survey data.)

Suppose that the two data sets are merged to identify the collection of potential matches for survey respondents in the administrative data set with $x_i = x = x_j^*$. Let n_x be the number of such subjects in the survey data base and the corresponding collection of survey respondents is $R_x = \{I_i, i = 1, 2, ..., n_x \mid x_i = x\}$. The collection of potential matches to these respondents is $R_x^* = \{I_j^*, j = 1, 2, ..., N_x \mid x_j^* = x\}$. The risk can be assessed in terms of the probability of a match, $p_{x,ij} = \Pr(I_i = I_j^* \mid x_i = x_j^* = x)$. Assuming that $R_x \subseteq R_x^*$ and no further information is available to refine the matches, this probability can be expressed as n_x / N_x.

For several values of *x*, the number of matches, N_x, may be large and the ratio n_x / N_x may be quite small. The risk may be high when both n_x and N_x are relatively small. For sample and population uniques (or for that matter any $n_x = N_x$), the probability of disclosure is 1. The risk of disclosure can be enumerated by studying the relationship between n_x and N_x as a function of *x*.

While deriving the probability of disclosure, we assumed that $R_x \subseteq R_x^*$. This assumption may be violated for at least two situations. First, the administrative data base may not cover the target population used by the survey. We may have respondents in the survey who are not part of the administrative database. In this case, the risk of disclosure is 0 for

m_x respondents who are not in the administrative database and the risk is $(n_x - m_x)/N_x$ for the remaining respondents who are in the survey.

The second reason leading to the possible violation of assumption $R_x \subseteq R_x^*$ is false matches due to measurement error. Many respondents may not be among a set of potential matches because the key variables used to match the two data sets may differ on the same individual. For example, suppose the gender of the female survey respondent is incorrectly stated as a male in the administrative data set and gender is one of the matching variables then this survey respondent is not in R_x^* and, hence, the probability of disclosure is 0.

In order to understand the risk of disclosure in a practical contexts, it is important to study the underlying "coverage probability," $r_{i,x} = \Pr(I_i \in R_x^* \mid x_i = x)$, for individual *i* with the specific demographic matching variables value *x*. Even when we obtain very small values of n_x and N_x, the risk of actual disclosure may be small because the coverage probability is also very small or even 0. At this point, we are only interested in the risk of exposing the true identity of survey respondent I_i to the intruder and are not concerned with the potential wrong identification. We discuss this issue more fully in Section 4.

3. Two Experiments

A large national population database with approximately 120 million records was leased from a commercial vendor. These data were complied from a number of public and private sources. Information is provided for up to 6 persons listed at a single address. However, in most cases data is available for only two persons. Included in the file are identifiers, names and addresses, and several key demographic variables. In addition to demographic data, the database also includes a number of indicators of wealth, purchasing behavior, and leisure and profession activities. The database is primarily intended for marketing and market research.

Table 1 lists the eight key population demographic variables used as matching variables in our experiments. Table 2 gives the missing data percentages on the eight variables in the population data base (after collapsing categories). The missing data percentages are not that substantial but it is likely that some imputations may have been performed.

We used several national surveys with the same demographic variables but with varying survey variables of interest such as health, income, drug use. All these surveys used area probability sample design. Data from these surveys were not subjected to any treatment to limit disclosure or the data prior to any such treatment were made available for this project. The total sample size across these surveys exceeded 200,000. We conducted two experiments. In the first experiment, the goal was to assess risk of disclosure in the surveys conditional on the population database. In the second, the goal was to assess the extent of measurement and coverage errors in the population database. We are not

identifying the surveys or the commercial database used in this experiment to protect confidentiality.

In the first experiment, we matched the survey respondents to the commercial database on the eight common variables to develop a list of potential matches for each respondent. That is, using the commercial database and each survey data we obtained n_x and N_x for each unique combination of variables listed in Table 1. Several combinations of large values of n_x and N_x were deemed to have lower risk of disclosure. We focused only on cells with $n_x \leq 5$ and $N_x \leq 25$. The respondent names and/or addresses from the survey data were then compared with the names and/or addresses in the population data. This experiment was repeated for all the national surveys included in the study. In none of the surveys used in this experiment, were we able to correctly identify the respondents.

None of the surveys we considered currently release geographical details beyond the Census regions as part of the survey database. However, for a few selected surveys, we were able to simulate what would have happened, if the survey had decided to release, for example, State or even lower level geographical identifiers. We had varying success in correctly identifying correct zipcode, city or state. However, with the release of such additional geographical details, additional information may become available to the intruder beyond the population database used in this experiment.

With this encouraging result (or discouraging depending upon one's perspective), we conducted the second experiment, wherein we matched addresses from one of our

surveys to addresses in the population database. Approximately, 74% of the addresses matched between the two sources. That is, there is 26% chance that the respondent in a survey may not be in the population database.

We then compared all eight demographic characteristics from the two data sources. For each characteristic, we created a binary indicator variable taking the value 1, if the characteristic matched a person in the population database to the person in the survey at the same address, and 0 otherwise. We then computed an average for each binary indicator. The average for each characteristic is provided in Table 3. These probabilities can be interpreted as the marginal probability of a match on the characteristic between the two data bases. Clearly, there are considerable mismatches in the basic characteristics between the survey reported and those in commercial database. There is more agreement in gender, race/ethnicity and marital status but none approach complete accuracy that is needed to precisely identify the respondent.

Table 4 provides the frequency distribution of the sum of the eight binary variables on each individual. This sum represents the number of characteristics that matched between survey and population databases. For example, approximately 0.01% of the survey cases matched to all 8 characteristics of Person 1 in the population database.

Two of the eight characteristics, the number of children in the household and household size were deemed to be least matching between the two data sources. We considered several alternatives: (1) Dropping household size and recoding the number children in the house hold to presence or absence of children; (2) Dropping both the variables from the list of characteristics; and (3) collapsing some categories. Obviously, all these steps increase values of both n_x and N_x. We also tried using a 2-year window while matching on the age to see whether it improved matching of individuals in the two data sources. None of these alternatives improved our ability to identify the respondents in the surveys.

4. Conclusions and Discussions

There are many statistical models for estimating disclosure risk based on the distribution of key variables (usually demographic). However, in this paper we are reporting on two experiments where we attempted to identify the respondents in national surveys using a population database. The first experiment indicated that a chance of respondent identification was rather low or nonexistent. Deeper investigation (experiment two) lead to the conclusion that the characteristics available to the intruder may not match very well with the characteristics reported by the respondent to the survey data collecting agency. That is, the measurement error or disagreement between the data sources acts as protection against disclosure.

It is important, however, to consider the distinction between Type I and Type II intruder problems while assessing disclosure risk. In Type I intruder there is no need for any external database to assess the risk of disclosure. Any sample unique based on certain

characteristics of the respondent that may be known by the intruder poses a risk. Here also the risk will be lower if the information provided by the respondent differs from the intruder's knowledge. For example, slightly differing view between the respondent and the intruder on race may throw the intruder off the track. Our focus in this paper was more on Type II intruders.

There are number of limitations in this research. We considered a broad spectrum of surveys in this experiment. But by no means can these surveys be considered exhaustive. We used the published sample design descriptions to glean geography as much as possible. Usually, these are used to define primary sampling units and are often masked. However, assumed intruder knowledge in this experiment could differ from other situations.

The commercial database we used as the population database may be limited. We arrived at this database after considerable research and this data source is used by many financial and marketing analysts through out the country. The coverage seems to be quite good. Some of the non matches were due to problems on the survey side. That is, the address descriptions in the survey were poor. For a few addresses in the commercial database, we checked the addresses using the Internet White pages and mostly they were correct. Further experiments with surveys that do release more geographical detail other than State or Region may be needed to fully assess the risk.

This research indicates that the potential risk of identifying a respondent, within the scope of the experiments outlined in this article, may be low and we may be overstating the risk of disclosure. The statistical steps such as data swapping or masking or cell suppression that agencies undertake to protect confidentiality may be overly conservative to protect against Type II intruder. Identification of a random respondent in a survey by an intruder will be quite rare. However, statistical disclosure limitation methods still may be needed to protect against Type I intruder. These techniques have to be evaluated in the context of protecting only vulnerable subjects from the intruder who knows that the subject participated in the survey. This may be a parent or a spouse of a respondent. Further experiment is needed to assess the actual risk posed by Type I intruder.

Acknowledgment

The research was partially supported by NIH grant PO1 HD045753 and NSF grant SES-0427889

References

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

Duncan, George, and Diane Lambert. 1989. "The Risk of Disclosure for Microdata". *Journal of Business Economics*, 7: 207-217

Elamir, Elsayed A.H., and Chris J. Skinner. 2006. "Record Level Measures of Disclosure Risk for Survey Microdata." *Journal of Official Statistics,* 22: 525-539

Fuller, Wayne A. 1993. "Masking Procedures for Microdata Disclosure Limitations." *Journal of Official Statistics*, 9: 383-406

Fienberg , Stephen E., and Udia E. Makov. 1998. "Confidentiality, Uniqueness, and Disclosure Limitation for Categorical Data." *Journal of Official Statistics*, 14: 385-397

Lambert, Diane. 1993. "Measures of Disclosure Risk and Harm." *Journal of Official Statistics*, 9: 313-331

Liu, Fang . 2003. "Bayesian Methods for Statistical Disclosure Control in Microdata." Ph.D. Dissertation, University of Michigan

Reiter, Jerome P. 2005. "Estimating Risks of Identification Disclosure in Microdata." *Journal of the American Statistical Association,* 100: 1103-1112

Skinner, Chris J., and Michael J. Elliot. 2002. "A Measure of Disclosure Risk for Microdata." Journal of the Royal Statistical Society, Ser B. 64: 855-867

Skinner, Chris J., and D. J. Holmes. 1998. "Estimating the Re-Identification Risk Per Record in Microdata." *Journal of Official Statistics*, 14: 361-372

Sweeney, Latanya. 1998. "Three computational systems for disclosing medical data in the year 1999. Medinfo, 9:1124-1129..

Sweeney, Latanya. 2000. "Foundations of Privacy Protection from a Computer Science Perspective." Proceedings, Joint Statistical Meeting, AAAS, Indianapolis, IN

Table 1: Population database demographic characteristics used in the identification experiment

Variable	**Level of details in the population database**
Gender	Male Female
Age	Years (19-87)
Race/Ethnicity	African-American Asian Mediterranean Native American Scandinavian Polynesian Middle Eastern Jewish Western European Eastern European Other
Marital Status	Married (includes Divorced/Widowed) Never Married
Education	Less than High School High School Diploma Some College Bachelors Degree Graduate Degree
Household Income (000)	1-14 15-24 25-34 35-49 50-74 75-99 100-124 125-149 150-174 175-199 200-249 250
Number of Children in Household	0-7
Household Size	1-9 or more

Table 2: Percent of demographic characteristics missing in population database by person

Variable	Definition	Percent Missing	
		Person 1 (n=121 million)	Person 2 (n=55 million)
Gender	Male/Female	15.1	11.2
Age	<19, 19-23, 24-29, 30-34 35-49, 50-64, 65 or more	9.7	6.8
Race/Ethnicity	White, Black, Hispanic, Other	17.7	14.7
Marital Status	Married, Never Married	31.0	34.6
Education	Less than High School, High School, Some College, College	9.8	6.9
Household Income (000)	<50, 50-74, 75 or more	9.7	< 0.001
Number of Children in Household	0, 1, 2 , 3 or more	9.7	--
Household Size	1, 2, 3, 4, 5, 6 or more	9.7	--

Table 3: Percent of demographic characteristics matched between survey and population databases by person at matched address

Variable	Definition	Percent Matched	
		Person 1	Person 2
Gender	Male/Female	48.2	64.1
Age	<19, 19-23, 24-29, 30-34 35-49, 50-64, 65 or more	31.0	36.6
Race/Ethnicity	White, Black, Hispanic, Other	70.1	77.1
Marital Status	Married, Never Married	62.8	67.9
Education	Less than High School, High School, Some College, College	20.4	21.7
Household Income (000)	<50, 50-74, 75 or more	52.2	44.6
Number of Children in Household	0, 1, 2 , 3 or more	51.6	44.6
Household Size	1, 2, 3, 4, 5, 6 or more	28.1	25.6

Table 4: Frequency distribution of total number of demographic characteristics matched between survey and population databases by person at matched address

Number of Characteristics	Percent of Characteristics Matched	
	Person 1	Person 2
0	6.7	14.0
1	9.8	30.3
2	17.2	22.6
3	27.0	13.6
4	22.2	105
5	12.2	5.6
6	3.9	2.3
7	0.9	0.7
8	0.01	0.2