Journal of Official Statistics, Vol 13, No. 1, 1997, 75–89

A Bayesian Approach to Data Disclosure: Optimal Intruder Behavior for Continuous Data

Stephen E. Fienberg,[1] *Udi E. Makov,*[2] *and Ashish P. Sanil*[3]

In this article we develop an approach to data disclosure in survey settings by adopting a probabilistic definition of disclosure due to Dalenius. Our approach is based on the principle that a data collection agency must consider disclosure from the perspective of an intruder in order to efficiently evaluate data disclosure limitation procedures. The probabilistic definition and our attempt to study optimal intruder behavior lead naturally to a Bayesian formulation. We apply the methods in a small-scale simulation study using data adapted from an actual survey conducted by the Institute for Social Research at York University.

Key words: Confidentiality; disclosure limitation; inferential disclosure; identity disclosure; measurement error.

1. Introduction

There has been a longstanding government interest and concern in the United States and elsewhere over the confidentiality of statistical data, especially as gathered in sample surveys and censuses. For example, the U.S. Bureau of the Census operates under Title 13 of the U.S. Code, virtually from its inception in 1929. Such legal guarantees of confidentiality are not only a reflection of the public concerns regarding disclosure but also of the agencies' desire for high quality data. Even in the absence of legal restrictions on access to data, statistical agencies and survey researchers have always been concerned about the need to preserve the confidentiality of respondents in order to ensure the quality of the data provided, and these concerns have been heightened by the decline in response rates for censuses and surveys over the past two decades (e.g., see Panel on Privacy and Confidentiality as Factors in Survey Response 1979; Fienberg 1993–1994).

At the same time government agencies have an obligation to report their data widely and thus they recognize the need for some balance between strict confidentiality (however it is to be interpreted) and the benefits derived from the release of statistical information. To

[1] Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A.
[2] Department of Statistics, Haifa University, Haifa, 31905, Israel.
[3] Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A.

Acknowledgments: The preparation of this article was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada to the first author at York University, North York, Ontario, Canada, and in part by the U.S. Bureau of the Census. We are especially grateful to the Institute for Social Research at York University which provided us with the data used in Section 4, and to Sangeeta Agrawal for assistance with the structuring of the database and a variety of initial calculations. We received valuable comments and suggestions on earlier drafts of the manuscript from staff at the U.S. Bureau of the Census, an Associate Editor and two referees, and these have led to substantial changes and improvements to the article. A preliminary version of this article was presented at the Second International Seminar on Statistical Confidentiality, Luxembourg, November 1994, and has been published as Fienberg, Makov, and Sanil (1995) under the same title.

judge the balancing of disclosure risk and the benefits to be derived from the access to data the statistician requires a technical interpretation for disclosure. This article offers a systematic attempt to examine these issues.

Various authors have attempted to provide a precise technical definition of the concept of disclosure. For example, Fellegi (1972) suggests that disclosure requires both the recognition of an individual member of a population included in a data release and learning something about that individual. In the context of sample surveys, the first part of the definition would mean that someone could actually identify a sample member on the basis of the data released without knowing a priori that the individual was a member of the sample. The second part means that just identifying someone as a sample member by a unique set of characteristics is not a disclosure without there being additional characteristics which are then identifiable. Fellegi goes on to discuss the notions of direct and residual disclosure.

In this article, we develop a Bayesian approach to the issue of data disclosure that builds upon a probabilistic approach first suggested by Dalenius (1977) who offered the following definition: *If the release of certain statistical information make it possible to determine a particular value relating to a known individual more accurately than is possible without access to that data, then a disclosure has taken place*. Because almost any data release provides some information about the individuals whose data are included, total avoidance of disclosure is impossible. Thus we are left with the notion of controlling or limiting disclosure. Duncan and Lambert (1989) describe this notion as inferential disclosure and contrast it with other notions of disclosure proposed in the literature. Note that we can have a disclosure according to this definition for someone in the population who has not actually provided data. Thus a disclosure does not always produce a breach of confidentiality.

Our work is guided by the principle that a data collection agency must consider disclosure from the perspective of an intruder in order to efficiently evaluate disclosure limitation procedures (c.f., Duncan and Lambert 1989; Lambert 1993). Others have also studied aspects of intruder behavior, most notably the empirical study by Paass (1988), Blien, Wirth, and Müller (1992), Fuller (1993), and Skinner, et al. (1994). Paass (1988) studied empirically the ability of an intruder to match two large files using a non-Bayesian discriminant analysis and he claims a substantial rate of success, even in the presence of added noise, but his results are not supported by another empirical study by Blien, Wirth, and Müller (1992). Lambert (1993) suggests a Bayesian approach to disclosure identification, using a logistic regression model for discrimination purposes, but she makes no attempt to optimize intruder behavior nor to allow explicitly for a mechanism that might prevent an intruder from making a match, i.e., measurement error. On the other hand, Fuller (1993) studies the role of measurement error as part of a masking process intended to foil an intruder's attempts at identification. Our work in this article shares Bayesian features (but not the linear logistic regression model) with Lambert and aspects of measurement error with Fuller. We do not suggest that an investigator necessarily must adopt Bayesian methodology for the sake of disclosure. We firmly believe, however, that Bayesian methodology is better suited for this purpose, in part because it allows for the direct incorporation of prior judgements and produces information on the probability of accurate identification. Consequently, we believe that Bayesian analysis provides a more accurate picture of data base vulnerability.

The risk of disclosure for population data, as in a census, is clearly greater than for exactly the same kinds of data releases for sample data. Indeed, the real aim of an organization addressing the issue of confidentiality is the exercise of disclosure control and the acceptability of disclosure risk associated with various kinds of data in different situations (see the discussion in Skinner et al. 1994). There is general agreement that, after specific identifiers, geographic information poses one of the greatest risks for disclosure and statistical agencies and survey organizations have developed rules on the suppression of detailed geographic information (Greenberg and Zayatz 1992). Recent work at the U.S. Census Bureau attempts to address this analytical concern through the creation of microdata files of ''contextual'' variables that present reduced disclosure risk (Saalfeld, Zayatz, and Hoel 1992). Duncan and Lambert (1986) illustrate how this probabilistic approach can be applied to provide a justification for various ad hoc rules proposed or actually used to limit disclosure.

Lambert (1993) argues that one can make an assessment of the ad hoc rules currently in use only when we have a working model for the behavior of the intruder. We follow this approach in Sections 2 and 3, as we develop a model for optimal identification of individual records by an intruder leading to the potential disclosure of information in an agency's data base. In Section 4, we present a small scale case study using data from a survey conducted by the Institute for Social Research at York University. The use of our model for the assessment of disclosure avoidance procedures still requires careful investigation.

2. Basic Disclosure Model

We assume the intruder possesses verified information, $\mathbf{x}$, on several individuals. He or she attempts to obtain additional information on one or more individuals by examining the released data, $\mathbf{y}$, and identifying the released information of these individuals with his or her records in $\mathbf{x}$. Our Bayesian treatment of this problem assumes: (i) a model describing the data with unknown parameters to which exchangeable priors are assigned; (ii) a probabilistic mechanism which introduces bias into the responses; (iii) a probabilistic mechanism which generates errors in the data base. Finally, we have the realistic assumption that there is a nonzero probability that the released data, $\mathbf{y}$, does not contain any information on the individual(s) concerned. The framework here is related to that used by Duncan and Lambert (1989).

In the simplest situation, the data available to the ''intruder'' consists of data on an individual who is at the center of the investigation ($\mathbf{x}_0$), where $\mathbf{x} = \{x_{01}, x_{02}, \ldots, x_{0k}\}$ is a k-attribute vector of observations collected on that individual. These k variables in the investigator's data are sometimes referred to as *key variables* which are also available in the data released by the agency (Bethlehem, Keller, and Pannekoek 1990; Skinner et al. 1994).

Assumption A1: The intruder saves no effort in verifying his or her data $\mathbf{x}_0$ which are therefore totally accurate.

While in most real-world situations the intruder's knowledge about the values of his or her data elements may contain considerable uncertainty, Assumption A1 allows us to put an upper bound on the posterior probability of correct identification.

The agency records data on N individuals which, for the purpose of simplification, we assume to constitute an entire population. We denote these data by $\mathbf{y}^{(N)} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ where $\mathbf{y}_i = \{\mathbf{y}_{i0}, y_{i1}, \ldots, y_{iq}\}$ is a $q + u$-attribute vector of observations recorded on the ith individual and where $\mathbf{y}_{i0}$ is a u-vector of ''identifying'' attributes (such as social security number) not released to the public. Clearly $\mathbf{y}_i \neq \mathbf{y}_j$ for $i \neq j$, as $\mathbf{y}_{i0} \neq \mathbf{y}_{j0}$. The agency releases the censored records of a subset of n individuals, $1 \leqslant n \leqslant N$, which we renumber from 1 to n (typically $1 \ll n$). We denote the released records by $\mathbf{z}^{(n)} = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$, where $\mathbf{z}_i = \{z_{i1}, \ldots, z_{iq}\}$ and assume that the attributes are arranged such that z_{ij} corresponds to x_{0j} for $j = 1, \ldots, k < q$.

The variables could be either discrete and/or continuous, but in this article we shall treat all variables as continuous, in order to avoid a focus on ''uniqueness'' (Bethlehem et al. 1990; Skinner et al. 1994). For continuous variables all observations are unique. While much survey and census data involve categorical variables or variables that are made such (e.g., the use of intervals for continuous quantities such as age or income), there are a number of practical survey contexts where interest focuses on continuous variables. For example, Kim and Winkler (1995) describe work done at the U.S. Bureau of the Census linking data from March 1990 supplement Current Population Survey to income data for the Internal Revenue Service Form 1040. The primary variables of interest to users of the merged data files are components of income that are all essentially continuous and preserving the confidentiality of the merged files has been a major preoccupation of those who worked on this project. Kim and Winkler (1995) use matrix masking techniques for the income variables in this dataset assuming an underlying multivariable normal distribution.

Even though we consider the case of continuous variables, an intruder may have only approximate information about some of them. By assuming exact knowledge on the part of the intruder, we are greatly simplifying the modeling problem and also getting an upper bound on the probability of correct matches in the more complex and realistic setting.

Assumption A2: The distribution of the attributes amongst all individuals is given by $f(\mathbf{z}^{(n)}|\boldsymbol{\mu})$, where $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_q\}$ is a vector of parameters. We assume independence between individuals, i.e., $f(\mathbf{z}^{(n)}|\boldsymbol{\mu}) = \Pi_i f(\mathbf{z}_i|\boldsymbol{\mu})$. We further assume that, a priori, $\boldsymbol{\mu} \sim t(\boldsymbol{\mu})$ (where $f(\cdot)$ and $t(\cdot)$ denote probability density functions).

The independence assumption is certainly not reflective of real-world settings and we have made it largely for convenience. Elsewhere we plan to relax this assumption to allow for interesting covariance structures.

We now define $\mathcal{J}$ be an indicator function, $\mathcal{J} = 1, 2, \ldots, n+1$, such that $\mathcal{J} = j$, (for $j = 1, \ldots, n$) if $\mathbf{x}_0$ is associated with the individual whose released record is given by $\mathbf{z}_j$, and $\mathcal{J} = n + 1$ if $\mathbf{x}_0$ is associated with an individual whose record has not been released. The aim of the investigator is to find the value of $\mathcal{J}$ (say j) since with such knowledge the confidential attributes $y_{j(k+1)}, \ldots, y_{jq}$ become known and identity disclosure is accomplished.

The task of identification is far from trivial, not only because the records corresponding to $\mathbf{x}_0$ are not necessarily in the released sample, but also because $\mathcal{J} = r \leqslant n$ does not imply that $z_{rj} = x_{0j}$. This is because the agency records, $\mathbf{z}^{(N)}$, can be accidentally corrupted or modified (intentionally or otherwise) by the individual supplying the information to the agency. While one can model a variety of noises and biases which are introduced into the data, we shall, for the sake of simplification, assume that disturbances are generated according to the following assumption:

Assumption A3: $x_{0j} = \theta_{ij} z_{ij} + \xi_j, \mathcal{J} = i; i = 1, \ldots, n+1; j = 1, \ldots, k.$

Here, θ_{ij} is an unknown, non-zero bias removing parameter associated with the jth attribute of the ith individual, and $\theta_{(n+1)j}$ is the parameter associated with any of the individuals whose records are not released by the agency. The random variable ξ_j follows a noise distribution, say $N(0, \sigma_j^2)$ with unknown variance, σ_j^2. Finally, we take the σ_j^2 to have distribution $\upsilon(\sigma_j^2)$, independent of θ_{ij}. Note that we assume here that the noise ξ_j varies across attributes but not across individuals. This assumption can be easily extended to account for changes in noise between individuals as well.

One may argue that rather than generating $\mathbf{x}$ from $\mathbf{z}$ via a ''bias removing parameter,'' we could generate $\mathbf{z}$ from $\mathbf{x}$ via a ''bias inducing parameter.'' The later formulation proved more cumbersome and thus we used the former.

Clearly if $\theta_{ij} = 1$ there is no bias and the ith respondent honestly reports the value of attribute j. However, due to additional noise (like typing errors etc.) modeled by the normal distribution, z_{ij} can still differ from x_{0j}.

We might argue that, in most cases, respondents provide honest (unbiased) answers ($\theta_{ij} = 1$) and that the degree of honesty depends both on the 'sensitivity' of a particular attribute j and on the subjective feelings of an individual i towards exposing the truthful information. We could go further and argue that there is a pattern of bias amongst the respondents in such a way that for an attribute generating bias (such as income) most people exhibit a bias of a similar nature (overstating/understating) though will differ in the extent of their departure from the truth. This leads us to:

Assumption A4: The θ_{ij} are exchangeable with respect to the index i and are distributed $g(\theta_{ij}|\varphi_j)$. The prior distribution of φ_j is given by $h(\varphi_j)$.

The problem faced by the intruder is to attempt to match his or her record $\mathbf{x}_0$ with one of those in the file released by the agency. What distinguishes this problem from the more basic problem of exact matching (e.g., see Subcommittee on Matching Techniques 1980) is the error structure embedded in the model outlined in this section. To cope with these errors, we now explore a Bayesian approach to identity disclosure.

3. Identity Disclosure and the Bayesian Approach

In attempting to establish identity disclosure, the intruder can use the posterior distribution of $\mathcal{J}$ given all available data, $P(\mathcal{J}|\mathbf{x}_0; \mathbf{z}^{(n)})$, and ''decide'' that the confidential records associated with $\mathbf{x}_0$ are in z_m, where m is the value of $\mathcal{J}$ for which the posterior distribution of $\mathcal{J}$ is maximized. We now discuss the evaluation of $P(\mathcal{J}|\mathbf{x}_0;\mathbf{z}^{(n)})$.

Using Bayes's Theorem we have

$$P(\mathcal{J}=i|\mathbf{x}_0;\mathbf{z}^{(n)}) \propto f(\mathbf{x}_0|\mathcal{J}=i;\mathbf{z}^{(n)})f(\mathcal{J}=i|\mathbf{z}^{(n)}), \qquad i=1,\ldots,n+1. \tag{1}$$

The second term on the right-hand side (r.h.s.) of (1) is simply the prior distribution of $\mathcal{J}$ since the observations contribute no extra information on $\mathcal{J}$. Clearly, in the absence of data on the individual at the center of investigation, we can take

$$f(\mathcal{J}=i|\mathbf{z}^{(n)}) = \begin{cases} 1/N & \text{for } i=1,\ldots,n \\ (N-n)/N & \text{for } i=n+1. \end{cases} \tag{2}$$

Note that as the difference between n and N increases we put less prior probability on individuals in the released data. The first term on the r.h.s. of expression (1) is the predictive distribution of $\mathbf{x}_0$, which is given, for $i=1,\ldots,n$, by

$$\begin{aligned} &f(\mathbf{x}_0|\mathcal{J}=i;\mathbf{z}^{(n)}) \\ &\quad = \prod_{j=1}^{k}\int_{\theta_{ij}}\int_{\sigma_j^2} f(x_{0j}|\theta_{ij};z_{ij};\sigma_j^2)f(\theta_{ij}|\mathbf{z}^{(n)};\mathcal{J}=i)f(\sigma_j^2|\mathbf{z}^{(n)};\mathcal{J}=i)d\theta_{ij}d\sigma_j^2 \end{aligned} \tag{3}$$

where the first term on the r.h.s of (3) is known from assumption A3.

For $i=n+1$, the value of z_{ij} is not released and therefore we need to integrate out the unknown realization of $z_{(n+1)j}$ by taking its expectation with respect to the predictive distribution $f(z_{(n+1)j}|\mathbf{z}^{(n)})$. Expression (3) now takes the form

$$\begin{aligned} &f(\mathbf{x}_0|\mathcal{J}=n+1;\mathbf{z}^{(n)}) \\ &\quad = \prod_{j}^{k}\int_{\theta_{(n+1)j}}\int_{\sigma_j^2}\int_{z_{(n+1)j}} f(x_{ij}|\theta_{(n+1)j};z_{(n+1)j};\sigma_j^2)f(\theta_{(n+1)j}|\mathbf{z}^{(n)};\mathcal{J}=n+1) \\ &\qquad * f(z_{(n+1)j}|\mathbf{z}^{(n)})f(\sigma_j^2|\mathbf{z}^{(n)};\mathcal{J}=n+1)\,d\theta_{(n+1)j}\,dz_{(n+1)j}\,d\sigma_j^2 \end{aligned} \tag{4}$$

where, by assumption A2,

$$f(z_{(n+1)j}|\mathbf{z}^{(n)}) = \int f(z_{n+1)j}|\boldsymbol{\mu})f(\boldsymbol{\mu}|\mathbf{z}^{(n)})\,d\boldsymbol{\mu} \tag{5}$$

and

$$f(\boldsymbol{\mu}|\mathbf{z}^{(n)}) \propto f(\mathbf{z}^{(n)}|\boldsymbol{\mu})t(\boldsymbol{\mu}) = \prod_{i=1}^{n} f(\mathbf{z}_i|\boldsymbol{\mu})t(\boldsymbol{\mu}). \tag{6}$$

4. A Case Study: Canadian Survey of Elites

4.1. The data

In this section we report preliminary simulation results using the formulation from Section 3. This case study uses data from the survey on Elite Canadian Decision-Makers collected by the Institute for Social Research at York University. This survey was conducted in 1981 using telephone interviews and there were 1,348 respondents, but many of these did not supply complete data. We have extracted data on 12 variables, each of which was measured on a 5-point scale:

Civil-liberties
$C1$ – Free speech is just not worth it
$C2$ – We have gone too far in pushing equal rights in this country
$C3$ – It is better to live in an orderly society than to allow people so much freedom
$C5$ – Free speech ought to be allowed for all political groups

Attitudes towards Jews
A15 – Most Jews don't care what happens to people who are not Jews
A18 – Jews are more willing than others to use shady practices to get ahead

Canada–U.S. relationship
$CUS1$ – Ensure independent Canada
$CUS5$ – Canada should have free trade with the U.S.A.
$CUS6$ – Canada's way of life is influenced strongly by U.S.A.
$CUS7$ – Canada benefits from U.S. investments

In addition, we have data on two approximately continuous variables:

Personal information
Income – Total family income before taxes (with top-coding at 80,000 USD)
Age – Based on year of birth

We transformed the original survey data as follows in order to create a database of approximately continuous variables:

A. We add categorical variables (all but income) to increase the number of levels. (When necessary we reversed the order of levels of a response to a question.) The new variables are defined as follows:

$$\text{Civil} = C1 + C2 + C3 + (8 - C5)$$

$$\text{Attitude} = A15 + A18$$

$$\text{Can/U.S.} = (5 - CUS1) + CUS5 + (5 - CUS6) + CUS7$$

After we removed cases with missing observations and two cases involving young children, we had a database consisting of 662 observations. (Extensions of the basic methodology to handle missing data are straightforward but beyond the scope of this article.)

B. In order to enhance continuity, we took the following measures:

Age: We added normal distributed variates, with 0 mean and variance 4 to all observations.

Income: We added uniform variates on the range of [0 USD–10,000 USD] to all incomes below 80,000 USD. Since all cases of incomes exceeding 80,000 USD were lumped together in the survey, we simulated their values by means of a $t(8)$ distribution. Drawing values from the upper 38% tail of $t(8)$, we evaluated the values of income as 60,000 USD $+ 25,000 * t(8)$.

Other variables: We added normal distributed variates, with 0 mean and variance 1/2 to the variables.

We assume that the agency releases information about all variables, except for Attitudes (towards Jews), which is unavailable to the intruder and is at the center of the intruder's investigation. We denote the released data by $\mathbf{z} = \{z_{ij}, i = 1, \ldots, 662, j = 1, \ldots, 4\}$. We further assume that the agency releases information on all records held ($n = N$), thus making the intruder's task easier than would be the case if only a sample were released.

We assume that the intruder's data are accurate and are related to $\mathbf{z}$ via the following transformation: $x_{0j} = z_{ij} * \theta_{ij} + \xi_j$, where θ_{ij} is a bias removing parameter normally distributed with mean 1 and variance φ_j, and ξ_j is normally distributed disturbance with 0 mean and variance σ_j^2. The following table provides the values of φ_j and σ_j^2 used in the study:

	φ_j	σ_j^2
Civil	0.1732	25
Can/U.S.	0.1732	25
Age	0.1732	9
Income (in 10,000's USD)	0.1732	4

Table 4.1. First 10 records of **x** *and* **z**

Age		Civil		Can/U.S.		Income (USD)	
x	**z**	**x**	**z**	**x**	**z**	**x**	**z**
44.607011	31.00364	24.803494	23.26688	0.2782396	7.230798	80351.260	86680.77
44.356572	58.36153	17.330712	17.76006	6.8772395	4.480846	95886.344	64127.42
35.260936	49.43488	10.930148	14.58399	12.3295743	7.632419	120247.969	88728.53
47.740238	40.87560	20.582654	16.54536	10.9438634	9.448964	106980.348	80348.58
32.257831	30.38650	21.430536	14.09269	16.5630154	10.828120	74050.109	76234.72
16.964057	21.51478	21.842566	14.62777	12.4165201	14.206017	105327.918	81986.36
43.319185	52.79831	10.552020	16.20542	2.3126535	5.881757	54703.225	73593.64
36.162886	42.55710	22.629968	21.26227	2.0760983	5.632542	39358.922	63209.81
31.119159	32.50106	15.738561	20.83966	8.4469488	14.843309	4466.606	42866.14
56.607847	82.03417	19.465088	18.70948	4.3898451	7.309711	102111.234	119271.81

In Table 4.1 we illustrate the impact of the process of the error on the data. The columns of this table present the intruder's accurate data, **x**, and the biased and corrupted released data, **z**, side by side. The marked difference between the two data sets, which in reality is unlikely to be so striking, clearly makes the task of identification by the intruder a difficult one.

4.2. Implementation of the Bayesian model

We assume a uniform prior distribution on all individuals in the released data, i.e., $f(\mathcal{J} = j|z^{(662)}) = 1/662$.

We attempt identification by evaluating equation (1), which we now modify following the particular assumptions on the distributions of θ_{ij}, and ξ_j. Taken together, $\xi_j \sim N(0, \sigma_j^2)$ and assumption A3 imply that $x_{ij}|\theta_{ij}, z_{ij}, \sigma_j^2 \sim N(\theta_{ij}z_{ij}, \sigma_j^2)$. The additional assumption, that $\theta_{ij}|\varphi_j \sim N(1, \varphi_j)$, implies that $x_{ij}|z_{ij}, \sigma_j^2, \varphi_j \sim N(z_{ij}; \sigma_j^2 + z_{ij}^2\varphi_j)$. Equation (3) now takes the form

$$\prod_{j}^{4} \int\int f(x_{0j}|\mathbf{z}^{(n)}; \sigma_j^2; \varphi_j; \mathcal{J} = i) f(\varphi_j|\mathbf{z}^{(n)}; \mathcal{J} = i) f(\sigma_j^2|\mathbf{z}^{(n)}; \mathcal{J} = i)\, d\varphi_j d\sigma_j^2. \tag{7}$$

Note that the $f(\psi_j|\mathbf{z}^{(n)}; \mathcal{J} = i)$ and $f(\sigma_j^2|\mathbf{z}^{(n)}; \mathcal{J} = i)$ cannot be evaluated unless the intruder possesses accurate information on additional individuals. Such a situation, however, would require entirely different modeling which we do not address here. Equation (7) therefore takes a simplified form

$$\prod_{j=1}^{4} \int\int f(x_{0j}|\mathbf{z}^{(n)}; \sigma_j^2; \varphi_j; \mathcal{J} = i) f(\varphi_j) f(\sigma_j^2)\, d\varphi_j\, d\sigma_j^2. \tag{8}$$

One possible identification rule is to choose that value of i which maximizes the left-hand side (l.h.s.) of (7). Of course, if the maximum posterior probability is small, a wise intruder should conclude that there is insufficient information to act as if identification has occurred.

In essence we employ Monte Carlo evaluation of the expectation in (8) using variates generated from $f(\varphi_j)$ and from $f(\sigma_j^2)$. In this study, we assumed that the prior distributions for σ_j^2 and φ_j, $\upsilon(\sigma_j^2)$ and $h(\varphi_j)$, respectively, are gamma distributions. We selected the values of the two hyperparameters of the gamma distributions in order to center the distribution at an ''intelligent guess'' of the mean value of the parameter. We guessed that the θ_j would range from 0.75 to 1.25 for all j. We took this range to be six times the (guessed) standard deviation of the θ_j's. This led to a guess of the central value or mean of the φ_j's. We guessed the mean values of the σ_j^2's in a similar manner. For instance, we took $\sqrt{\sigma_1^2}$, corresponding to *Age*, to be one-sixth of the range of the *Age* variable in the released dataset. In addition, we added a condition that the coefficient of variation (mean divided by standard deviation) is 20. We give the resulting α and β parameters we used for the gamma priors in the following table:

Table 4.2. Values of the parameters of the Gamma priors distributions used for φ and σ^2

	Attribute	α	β
φ	Age	400	5,000
	Civil	400	5,000
	Can/U.S.	400	5,000
	Income (in 10,000's USD)	400	5,000
σ^2	Age	400	3.5
	Civil	400	34
	Can/U.S.	400	63
	Income (in 10,000's USD)	400	110

4.3. Illustration of computations

To gain a better understanding of the performance of our Bayesian model for intruder behavior, we next conducted a complete simulation of the procedures for the complete set of $n = 662$ cases. We considered four different scenarios for the simulation:

- The released data contains no bias or noise (i.e., $\varphi_j = 0$ and $\sigma_j^2 = 0$ for all j).
- The released data contains only noise (i.e., $\varphi_j = 0$ for all j and σ_j^2 as given in Table (4.2)).
- The released data contains only bias (i.e., $\sigma_j^2 = 0$ for all j and φ_j as given in Table (4.2)).
- The released data contains both bias and noise (i.e., σ_j^2 and φ_j as given in Table (4.2)).

We took each individual in turn as the object of the intruder's efforts and carried out the calculations.

Table 4.3. Results for data without noise or bias

Rank	Number observed	Avg. prob. match (SD)	Top 10 cum. (SD)
1	319	0.017 (0.021)	0.113 (0.067)
2	144	0.009 (0.003)	0.078 (0.025)
3	106	0.007 (0.003)	0.066 (0.024)
4	41	0.006 (0.001)	0.06 (0.011)
5	34	0.006 (0.001)	0.058 (0.011)
6	8	0.005 (0.001)	0.051 (0.009)
7	3	0.005 (0.0001)	0.054 (0.001)
8	3	0.005 (0.001)	0.055 (0.008)
9	3	0.006 (0.0001)	0.056 (0.003)
10	1	0.007 –	0.074 –

We display the results in Table 4.3 through Table 4.6. The second column indicates the number of times the correct record in **z** was ranked 1st, 2nd, ..., 5th, or 11th and higher. The third column of the table gives the average probability of a correct match for each rank (with its standard deviation in parenthesis), and the fourth column gives the average sum of the probabilities of the highest ten probabilities of match for each rank (with its standard deviation in parenthesis).

In Table 4.3, where there is neither noise nor bias, we see that in 319 of 662 cases the intruder's posterior probability of a match was highest for the correct record; in 144 cases, the posterior probability of a match was second highest for the correct record; and so on. In this exact match case, the intruder does well but not perfectly, since he or she works to learn about the parameters of the distributions and not simply to match records. Moreover, the average value of the posterior probability associated with the highest ranked record is still modestly small, running roughly from 0.02 to 0.10, and the average value of the posterior probability associated with the 10 highest ranked records ran from about 0.25 to 0.40. There were no cases where the correct record ranked higher than 10th according to the intruder's posterior probability.

Table 4.4. Results for data with only noise

Rank	Number observed	Avg. prob. match (SD)	Top 10 cum. (SD)
1	42	0.045 (0.057)	0.204 (0.14)
2	36	0.023 (0.021)	0.162 (0.087)
3	29	0.015 (0.012)	0.177 (0.092)
4	26	0.015 (0.01)	0.146 (0.089)
5	18	0.015 (0.007)	0.16 (0.082)
6	22	0.014 (0.008)	0.145 (0.094)
7	6	0.014 (0.01)	0.193 (0.177)
8	13	0.01 (0.003)	0.119 (0.043)
9	12	0.009 (0.004)	0.111 (0.06)
10	10	0.012 (0.007)	0.176 (0.156)
>10	448	0.005 (0.003)	0.109 (0.058)

Table 4.5. Results for data with only bias

Rank	Number observed	Avg. prob. match (SD)	Top 10 cum. (SD)
1	295	0.018 (0.026)	0.116 (0.071)
2	141	0.009 (0.004)	0.081 (0.029)
3	75	0.008 (0.003)	0.073 (0.022)
4	46	0.006 (0.001)	0.06 (0.012)
5	33	0.006 (0.001)	0.062 (0.012)
6	21	0.007 (0.002)	0.067 (0.02)
7	13	0.005 (0.001)	0.054 (0.008)
8	11	0.005 (0.001)	0.056 (0.01)
9	6	0.006 (0.002)	0.061 (0.019)
10	2	0.006 (0.002)	0.062 (0.016)
>10	19	0.005 (0.0001)	0.054 (0.005)

Table 4.6. Results for data with both bias and noise

Rank	Number observed	Avg. prob. match (SD)	Top 10 cum. (SD)
1	43	0.044 (0.068)	0.193 (0.149)
2	34	0.026 (0.022)	0.179 (0.088)
3	34	0.02 (0.014)	0.168 (0.097)
4	24	0.014 (0.008)	0.133 (0.073)
5	16	0.02 (0.012)	0.2 (0.115)
6	15	0.013 (0.006)	0.139 (0.073)
7	16	0.014 (0.007)	0.167 (0.119)
8	13	0.014 (0.006)	0.184 (0.116)
9	12	0.011 (0.005)	0.139 (0.091)
10	4	0.009 (0.003)	0.104 (0.042)
>10	451	0.005 (0.003)	0.108 (0.057)

In Table 4.4, where there is noise but no bias, we see that in only 42 of 662 cases the intruder's posterior probability of a match was highest for the correct record; in 448 cases, the correct record was not even in the top ten in terms of posterior probability. The noise has led to a precipitous drop in the intruder's ability to infer a correct match. Things are not quite so difficult for the intruder when there is bias but no noise. From Table 4.5, we see that in 295 of 662 cases the intruder's posterior probability of a match was highest for the correct record, in only 19 cases the correct record was not in the top ten in terms of posterior probability.

Finally, in Table 4.6, where there is both noise and bias, we see that in only 43 of 662 cases the intruder's posterior probability of a match was highest for the correct record, while in 451 cases, the correct record was not in the top ten in terms of posterior probability. The bias leads to some degradation in matching over and above the noise, but not much.

From an examination of the top ten posterior choices of the intruder, we clearly see that, by applying the Bayesian model with carefully chosen priors, the intruder is able to perform moderately well, at least in terms of winnowing down the agency records to a small number of possible matches with moderately high probability. Even in the presence of moderate bias and noise in the released data the intruder has the correct match in the top five around 40% of the time.

5. Discussion

In this article, we have laid out a framework for a Bayesian approach to data disclosure in which we focus on the perspective of an intruder. In our framework, the intruder attempts to use microdata released by an agency in order to gain access to additional information on specific individuals for whom he or she already possesses information. The intruder attempts to ''break'' the confidentiality of the released data by comparing his or her individual-level data on specific individuals for a selection of variables with the values on those same variables for individuals in the released database. We chose to specialize our framework to the situation involving continuous data that are treatable as observations drawn from a normal distribution, for simplicity when carrying through the formal machinery of the Bayesian framework, and we made a number of simplifying assumptions regarding the noise and bias as well as the independence of variables.

The key to our framework is the role of error in the variables in the released database which, when combined with the continuous observations, precludes the use by the intruder of exact matches between data on individuals which he or she possesses and data in the released database. Thus, in the problem we have chosen to investigate in this article every individual in the released data file is ''unique'' but the data for no one can match exactly ''precise'' data possessed by an intruder. We believe that this framework and the key assumptions about the errors are realistic ones that are relevant to many problems of interest.

We then studied numerically the workings of our Bayesian model for intruder behavior using data based on observations from an actual survey (suitably modified). We found that, at least in the circumstances involving what we deemed to be the modest database error we chose to examine here, an intruder following the Bayesian model had great difficulty achieving matches with high posterior probability, and we would expect this probability to degrade further as n grows. In such uncertain settings an intruder would most often mis-identify which individual in the released data matched with the data in the intruder's possession. For example, suppose we had adopted some threshold for the intruder's posterior probability of a match and we had declared a match only when the maximum posterior probability exceeded that threshold. This would correspond to the specification of a loss function for the intruder, perhaps along the lines of those discussed in Duncan and Lambert (1989). Then we would have a negligible correct identification rate. These results should not be interpreted as indicating that there is no risk to the release of microdata files, but rather that error in the data plays a more substantial role than some previous investigators may have suggested.

In choosing the parameter settings for the database errors in our model, we explored a variety of values and decided, on somewhat subjective grounds, that we had modest levels of error present. Readers will have to judge for themselves whether they believe these to be modest error values.

Paass (1988) reports on a substantial empirical investigation involving the simulated matching of two extremely large data files, including the use of added noise, using a form of discriminant analysis and he concludes that there is a non-negligible rate of successful identification for target individuals randomly selected from the population. These results seem startlingly different from those suggested by our modest investigations. Yet, it is important to note that Blien, Wirth, and Müller (1992) also carried out a detailed empirical investigation, but using two actual data files (one of which involved data from the German microcensus). They found much higher levels of erroneous identification and substantially lower levels of correct identification. They concluded that the frequencies of error in the data which they observed serve as a ''natural barrier'' against intruders. These latter conclusions seem much closer in accord with the results reported here. We hope to shed further light on these widely discrepant conclusions through a more elaborate formal simulation using the Bayesian framework described in Sections 2 and 3.

This article is, in many senses, only a first consideration of a more general Bayesian approach to the problem of modeling optimal intruder behavior, and our effort has been necessarily specialized and limited in nature. In work in progress, we are attempting to remove the simplifying assumptions adopted in Section 2, and we are considering the more difficult case where the intruder matches more than one record at a time against the agency's released data, and where the released data are only a sample from the population and possibly even a sample of the data held by the agency. We expect an increase in the intruder's rate of success with an increase in the number of individuals whose record the intruder possesses. On the other hand, we expect an even more substantial decrease in his or her success rate with a drop in the sampling rate or the proportion of the data released by the agency. The proposed study will throw light on the actual effect of these factors and hence will provide some guidance as to how much of a decrease in the proportion of released data might be necessary to neutralize an intruder's increased resources. Further, we plan to investigate related models for intruder behavior when the variables under consideration are categorical or a mixture of continuous and categorical variables.

Once we have a systematic understanding of optimal intruder behavior in these varied circumstances, we hope to turn our attention to developing an exploration of the agencies response to an optimally performing intruder through the use of such techniques as matrix masking, and then finally assess whether an intruder can undo the agency's attempt to protect confidentiality. Our speculation from the calculations in the present article is that the addition of noise for continuous variables will be sufficient to protect confidentiality provided that the released file is moderately large in size. More noise will be required, however, if the number of key variables is increased.

6. References

Bethlehem, J.G., Keller, W.J., and Pannekoek, J. (1990). Disclosure Control of Microdata. Journal of the American Statistical Association, 85, 38–45.

Blien, U., Wirth, H., and Müller, M. (1992). Disclosure Risk for Microdata Stemming from Official Statistics. Statistica Neerlandica, 46, 69–82.

Dalenius, T. (1977). Towards a Methodology for Statistical Disclosure Control. Statistisk tidskrift, 5, 429–444.

Duncan, G.T. and Lambert, D. (1986). Disclosure-limited Data Dissemination (with Discussion). Journal of the American Statistical Association, 81, 10–28.

Duncan, G.T. and Lambert, D. (1989). The Risk of Disclosure for Microdata. Journal of Business and Economic Statistics, 7, 207–217.

Fellegi, I.P. (1972). On the Question of Statistical Confidentiality. Journal of the American Statistical Association, 67, 7–18.

Fienberg, S.E. (1993). Conflict Between the Needs for Access to Statistical Information and Demands for Confidentiality. Proceedings of International Seminar on Statistical Confidentiality. EUROSTAT: Luxembourg, 33–47.

Fienberg, S.E. (1994). Conflict Between the Needs for Access and Demands for Confidentiality. Journal of Official Statistics, 10, 115–132.

Fienberg, S.E., Makov, U.E., and Sanil, A.P. (1995). A Bayesian Approach to Data Disclosure. Optimal Intruder Behavior for Continuous Data. International Seminar on Statistical Confidentiality Proceedings. Official Publications of the European Communities: Luxembourg, 89–102.

Fuller, W. (1993). Masking Procedures for Microdata Disclosure Limitation. Journal of Official Statistics, 9, 383–406.

Greenberg, B.V. and Zayatz, L.V. (1992). Strategies of Measuring Risk in Public Use Microdata Files. Statistica Neerlandica, 46, 33–48.

Kim, J.J. and Winkler, W.E. (1995). Masking Microdata Files. Paper presented at the Joint Statistical Meetings, Orlando, Florida, August.

Lambert, D. (1993). Measures of Disclosure Risk and Harm. Journal of Official Statistics, 9, 313–332.

Panel on Privacy and Confidentiality as Factors in Survey Response (1979). Privacy and Confidentiality as Factors in Survey Response. Committee on National Statistics. National Academy of Sciences: Washington, DC.

Paass, G. (1988). Disclosure Risk and Disclosure Avoidance for Microdata. Journal of Business and Economic Statistics, 6, 487–500.

Saalfeld, A., Zayatz, L.V., and Hoel, E. (1992). Contextual Variables via Geographic Sorting: A Moving Averages Approach. Proceedings of the Section on Survey Research Methods, American Statistical Association: Washington, DC.

Skinner, C., Marsh, C., Openshaw, S., and Wymer, C. (1994). Disclosure Control for Census Microdata. Journal of Official Statistics, 10, 31–51.

Subcommittee on Matching Techniques (1980). Statistical Policy Working Paper 5: ''Report on Exact and Statistical Matching Techniques. Federal Committee on Statistical Methodology,'' Office of Federal Statistical Policy and Standards, U.S. Department of Commerce: Washington, DC.

Received October 1994
Revised July 1996