BAYESIAN STATISTICS 7, pp. 000–000
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)

Assessing the Risk of Disclosure of Confidential Categorical Data

ADRIAN DOBRA
National Institute of Statistical Sciences, USA
adobra@niss.org

STEPHEN E. FIENBERG
Department of Statistics, Carnegie Mellon University, USA
fienberg@stat.cmu.edu

MARIO TROTTINI
Departamento de Estadística e I.O., Universitat de València, Spain
Mario.Trottini@uv.es

SUMMARY

Disclosure limitation involves the application of statistical tools to limit the identification of information on individuals (and enterprises) included as part of statistical data bases such as censuses and sample surveys. We outline the major issues involved in assessing disclosure risk and assuring the protection of confidentiality for data bases, especially those in the form of multi-way contingency tables, and we present a Bayesian framework for thinking about such problems both from the perspective of an intruder and the agency trying to protect its data.

Keywords: CONTINGENCY TABLES; DATA UTILITY; DIRICHLET PRIOR; DISCLOSURE LIMITATION; INTRUDER BEHAVIOR; LOG-LINEAR MODELS.

1. INTRODUCTION

Maintaining the confidentiality of statistical data is essential if government agencies are to collect and publish high quality census and survey data. Typically agencies promise respondents that their data will be kept confidential and used for statistical purposes only. For example, Title 13, Section 9 of the United States Code prohibits the U.S. Census Bureau from publishing results in which an individual's or business' data can be identified. How can an agency comply with such legal strictures while at the same time provide public access to as much data as possible? This paper addresses this issue in the context of categorical data in the form of a cross-classification of counts.

Disclosure limitation is the process of protecting the confidentiality of statistical data. This paper focuses on *identity disclosure* where an intruder uses published statistical information to identify individual data provider. [For simplicity we set aside the issue of *attribute disclosure*, where an intruder learns that everyone in an identifiable group has a particular attribute.] Since virtually any form of data release contains some information about the individuals whose data are included in it, disclosure is not an all-or-none concept but rather a probabilistic one seen differently from the eyes of the agency protecting the data, the individuals providing the data,

and an "intruder" attempting to gain access to identifiable individual information (c.f., Lambert, 1993). In this sense, disclosure risk and the development of methods to limit disclosure are inherently Bayesian. For general introductions to some of the statistical aspects of confidentiality and disclosure limitation see Doyle, et al. (2001), Fienberg (1994), and Willenborg and De Waal (1996, 2001). Early Bayesian contributions to the literature on disclosure limitation include Duncan and Lambert (1986, 1989), and Rubin (1993).

Disclosure limitation procedures alter or limit the data to be released, e.g., by modifying or removing those characteristics that put confidential information at risk for disclosure. In the case of sample categorical data, a count of "1" can generate confidentiality concerns if that individual is also unique in the population. Much confidentiality research has focused on measures of risk that attempt to infer the probability that an individual is unique in the population given uniqueness in the sample (e.g., see Chen and Keller-McNulty, 1998, Fienberg and Makov, 1998, 2001, Skinner and Holmes, 1998, and Samuels, 1998). For simplicity, we focus on population tables of counts here and thus set aside this issue of making inferences from sample tables. But in either population or sample settings, small counts raise issues of disclosure risk.

In the next section we describe the identity disclosure problem in the context of a sequence of releases of marginal tables from a multi-way cross-classification. Then, in Section 3 and 4, we outline a Bayesian approach to the balancing of disclosure risk and data utility, apply it to the case of tabular categorical data, and derive some commonly used risk measures for the release of a sequence of marginal tables. In Section 5, we adopt the perspective of the intruder and consider updating distributions over the space of possible tables subject to margin constraints. We illustrate the methodology using a $2 \times 3 \times 3$ contingency table drawn from the 1990 U.S. decennial census. We conclude with a discussion of a number of unaddressed elements that need to be part of a full Bayesian approach to the problem.

2. DISCLOSURE LIMITATION FOR CATEGORICAL DATA

We think of the confidentiality problem as one involving three parties: a "statistical agency" that controls the data, "users" who wish to analyze all or perhaps subsets of the data, and an "intruder" who is attempting to identify one or more individuals in the data for some purpose.

Clearly any release of data from a database increases the information available about individuals in the database and thus increases in some sense the probability that an individual in the database will become identifiable. Harm to such an individual occurs when an intruder matches the identifiable record to an existing database and learns information about the individual that was not previously available. Following Fienberg, Makov, and Sanil (1997), we assume that the intruder acts as Bayesian updating his probabilities of identification of individuals in the database as more and more information becomes available. Further we assume that the agency acts in a Bayesian fashion and makes a trade-off between the utility of the data were it to be released to the users and the disclosure risk associated with that release.

Our goal here is to outline a statistical framework for the release of cross-classified categorical data in the form of a contingency table. We are thinking in terms of requests from users for (marginal) sub-tables involving a subset of the variables. Potential responses include the release of the requested sub-table, the release of an "altered" or "masked" sub-table, or perhaps a refusal to release the sub-table. Note that there is statistical information for an intruder that comes from a refusal, although we have yet to see a Bayesian analysis that takes such information into account. We think in terms of a public system so that once a subtable is released it is publicly available, and thus usable by an intruder. Clearly, the more subtables that are released, the more information we have about the full joint distribution of the cross-classifying variables.

The notion of data masking, introduced in Duncan and Pearson (1991) involves a trans-

formation to the data so that individual records are altered to make them less identifiable. For categorical data, when releases consist of marginal tables, the types of masks suggested in the literature include stochastic perturbations subject to the constraint that the transformed data are consistent with the released marginals (e.g., see Duncan, et al., 2001, and Fienberg, Makov and Steele, 1998).

How should the agency assess disclosure risk in this setting? What strategy should the intruder use to update his information about the individuals whose data are included in the full table? And, finally, given such choices, how should the agency respond to requests for specific marginal tables, given the set of tables already released?

We are unaware of any systematic and coherent statistical approach to the confidentiality problem as we have just outlined it, although Raghunathan and Rubin's (2001) multiple imputation strategy may provide a sensible Bayesian solution to it. Statistical agencies do in fact release subtables of very large contingency tables all of the time (e.g., the website for the U.S. Census Bureau's American Factfinder system releases selected three-way tables for various levels of geography: `http://factfinder.census.gov/`) and otherwise make judgments about the safety of releasing microdata files from sample surveys, the judgments about the "safety" of such data releases is ad hoc at best. Recent efforts to study the release of margins of contingency tables have focused on the role of bounds on cell entries that result (e.g., see Dobra and Fienberg, 2000, 2001, and Dobra, et al., 2002), and on perturbations of data based on "exact" distributions for contingency tables under log-linear models given marginals corresponding to minimal sufficient statistics (e.g., see Diaconis and Sturmfels, 1998, and Fienberg, Makov, and Steele, 1998). This work offers a starting point for the present paper in which we attempt to outline some of the elements of a Bayesian approach.

3. A GENERAL FRAMEWORK FOR ASSESSING DISCLOSURE RISK

Let $\boldsymbol{f}$ represent the original data and $\mathcal{D}$ a set of candidate data *masks* or transformations of the data, typically stochastic in nature. Here we outline a general Bayesian framework, based on Trottini (2001) and Trottini and Fienberg (2002), to answer the question: "Which mask should the agency select?" In Section 4, we apply the framework to tabular categorical data.

The evaluation of a generic data mask $\tilde{\boldsymbol{f}}$ depends on the extent to which its release is beneficial for the users, (*data utility* of $\tilde{\boldsymbol{f}}$) and the extent to which its release can harm the agency or the data providers (*disclosure risk* of $\tilde{\boldsymbol{f}}$). For simplicity, we assume that there are only two users of the data: an *intruder* (I) who wants to "undo" the candidate mask $\tilde{\boldsymbol{f}}$ to disclose confidential information about the data provider, and a *scientist* (S) who wants to use the released data to infer some general feature of the population underlying $\tilde{\boldsymbol{f}}$. We denote the intruder's target by Θ_I and the scientist's target by Θ_S. We assume that user h ($h = I, S$) incurs a loss $L_h(e, \Phi_h)$, by using the estimate e for his target Θ_h, which depends on an unknown "state of the world" Φ_h. In most cases of interest, $\Phi_h = \Theta_h$. We denote by $\pi_h(\cdot)$, and $\pi_h(\cdot \mid \tilde{\boldsymbol{f}})$ the users prior and posterior distributions for Φ_h, $h = I, S$.

We assume that both S and I act in accord with the expected loss principle, i.e. they estimate their target values by

$$\hat{\theta}_h = argmin_a \int L_h(a; \phi_h)\pi_h(\phi_h \mid \tilde{\boldsymbol{f}})d\phi_h, \quad h = I, S.$$

Following DeGroot (1962), we define user h's uncertainty about the true value of the target as the expected loss associated with the optimal estimate of the target,

$$U_h(\tilde{\boldsymbol{f}}) = \int L_h(\hat{\theta}_h; \phi_h)\pi_h(\phi_h \mid \tilde{\boldsymbol{f}})d\phi_h, \quad h = I, S.$$

We assume that the user stops trying to estimate Θ_h if his uncertainty is very large, in accord with the following decision rule:

User h's decision rule: *For a fixed threshold t_h, if $U_h(\tilde{\boldsymbol{f}}) \leq t_h$ then h takes action a_{1h} and estimates Θ_h by $\hat{\theta}_h$. If $U_h(\tilde{\boldsymbol{f}}) > t_h$ then h takes action a_{0h} and stops trying to estimate Θ_h.*

We assume that the loss that the agency incurs when user h takes action A_h, ($A_h \in \{a_{1h}, a_{0h}\}$) depends on an unknown state of the world $\Phi_A^{(h)}$ and is denoted by $L_A^{(h)}(\cdot,\cdot)$, $h = S, I$. Thus, $L_A^{(I)}(A_I, \phi_A^{(I)})$ quantifies, from the agency's perspective, the harm that the intruder's action A_I produces to the agency and the data providers when $\Phi_A^{(I)} = \phi_A^{(I)}$. In most of the cases it will be $\Phi_A^{(I)} = \Phi_I$. Similarly, $L_S^{(A)}(A_S, \phi_A^{(S)})$ quantifies, from the agency's perspective, the loss that the agency and the scientist incur if scientist takes action A_S and $\Phi_A^{(S)} = \phi_A^{(S)}$. In most of the cases this is just the loss that the scientist incurs by taking action A_S when $\Phi_A^{(S)} = \phi_h^{(S)}$, i.e., $\phi_A^{(S)} = \phi_S$ and $L_A^{(S)}(a_{1S}, \phi_S) = L_S(a_{1S}, \phi_S)$. We denote by $\pi_A^{(h)}(\cdot)$ and $\pi_A^{(h)}(\cdot \mid \tilde{\boldsymbol{f}})$ the agency's prior and posterior distribution for $\Phi_A^{(h)}$.

We assume that the agency treats A_h and the "states of the world" $\Phi_A^{(h)}$ as random variables, and we propose to measure *disclosure risk*, DR, and *data utility*, DU, averaging losses with respect to the agency's joint posterior distribution for A_h and $\Phi_A^{(h)}$ given the original data $\boldsymbol{f}$,

$$DR(\tilde{\boldsymbol{f}}) = E_{A_I, \Phi_A^{(I)} \mid \boldsymbol{f}}\{L_I^{(A)}(A_I, \phi_A^{(I)})\}, \quad DU(\tilde{\boldsymbol{f}}) = -[E_{A_S, \Phi_A^{(S)} \mid \boldsymbol{f}}\{L_S^{(A)}(A_S, \phi_A^{(S)})\}].$$

We assume that the users' targets as well as the users' priors, $\pi_h(\cdot)$, and the users' loss functions, $L_h(\cdot,\cdot)$, are known to the agency. This implies that the agency knows the users' posterior distributions, users' optimal estimate of Θ_h, $\hat{\theta}_h$, and users' uncertainty, U_h, $h = S, I$. We make this assumption largely for convenience and extensions to classes of targets, priors, and loss functions are possible.

We further assume that the users' thresholds, t_h, are fixed but unknown to the agency, which thus treats them as random variables independent of the state of the world $\Phi_A^{(h)}$. The independence assumption is reasonable if user h fixes his threshold on the basis of what he knows about Θ_h but never on the basis of the agency's knowledge of $\Phi_A^{(h)}$. It follows that the agency's posterior distribution for A_h, $\{\Pr(a_{0h} \mid \boldsymbol{f}), \Pr(a_{1h} \mid \boldsymbol{f})\}$, depends only on the agency's distributions, $\pi_{T_h}(\cdot)$, for the users' thresholds and we can rewrite the disclosure risk and data utility as:

$$DR(\tilde{\boldsymbol{f}}) = \sum_{j \in \{1,0\}} \Pr(a_{jI} \mid \boldsymbol{f}) \cdot E_{\Phi_A^{(I)} \mid \boldsymbol{f}}\{L_A^{(I)}(a_{jI}, \phi_A^{(I)})\}, \tag{1}$$

$$DU(\tilde{\boldsymbol{f}}) = -\sum_{j \in \{1,0\}} \Pr(a_{1S} \mid \boldsymbol{f}) \cdot E_{\Phi_A^{(S)} \mid \boldsymbol{f}}\{L_A^{(S)}(a_{1S}, \phi_A^{(S)})\}. \tag{2}$$

In most of the cases the agency does not know the users' target but can only identify classes $\mathcal{Z}_h$ of possible targets, i.e., $\Theta_h \in \mathcal{Z}_h = \{\Theta_h(1), \ldots, \Theta_h(r_h)\}$, and we can average (1) and (2) with respect to the probability that $\Theta_h = \Theta_h(j)$.

3.1. *The Utility-Risk Trade-off*

The most common criterion for the choice of the best mask in $\mathcal{D}$ consists of selecting the mask $\tilde{\boldsymbol{f}}$ that maximizes data utility subject to an upper bound for disclosure risk (Willenborg and de Waal, 2001, Duncan, Keller-McNulty, and Stokes, 2001, Trottini and Fienberg 2002). The optimal mask is the solution of the optimization problem:

$$max\{DU(\tilde{\boldsymbol{f}}) : \tilde{\boldsymbol{f}} \in \mathcal{D}, \text{ and } DR(\tilde{\boldsymbol{f}}) \leq \alpha\}$$

where α is a threshold value for the maximum tolerable risk fixed by the statistical agency. Defining an optimality criterion corresponds to specifying suitable measures of disclosure risk and data utility. We believe that the framework outlined in section 3 is the natural tool to define such measures. Once we have specified the users' targets, the information available about these targets prior to the release of the data, the estimation procedure used by the users, the consequences for the agency of users' actions, then (1) and (2) automatically provide measures of disclosure risk and data utility coherent with these inputs.

One might argue that all these elements are mostly unknown to the agency and, as a result, that our framework is difficult to implement, and that heuristic measures could do a better job. In fact, the uncertainty about inputs is a major strength of our approach, since our framework allows us to incorporate this uncertainty in a natural way. Heuristic measures are not assumptions free. Rather the assumptions simply are not stated (and therefore not understood). We have been able to use our framework to produce most of the measures of disclosure risk and data utility proposed in the literature of statistical confidentiality for suitable choices of the input values. This allows us to understand whether these measures are statistically sensible.

In the next section we apply the framework to tabular categorical data and, because of space limitations we focus only on measures of disclosure risk. Similar results hold for data utility.

4. DISCLOSURE RISK FOR TABULAR CATEGORICAL DATA

Suppose that a statistical agency records the value of k categorical variables for each individual in a given population and summarizes the result in a frequency table $\boldsymbol{f}$ with m cells (corresponding to the possible cross-classifications of the k variables). Let $\mathcal{I} = \{1, 2, \ldots, m\}$. We assume that the table total (population size), n, is known a priori to the users, who view the original table as a random variable, $\boldsymbol{F}$. Before the generic masked data $\tilde{\boldsymbol{f}}$ is released, users know that $\boldsymbol{F}$ takes values in the set $\mathcal{X}$ of all non-negative integer m-vectors adding to n

$$\boldsymbol{F} \in \mathcal{X} = \{(x_1, \ldots, x_m) : x_i \text{ is a non-negative integer and } \sum_{i=1}^{m} x_i = n\}.$$

We let $\mathcal{T}$ be the set of tables in $\mathcal{X}$ that are compatible with the candidate release $\tilde{\boldsymbol{f}}$ and by $M(\mathcal{X})$ and $M(\mathcal{T})$ the cardinality of $\mathcal{X}$ and $\mathcal{T}$ respectively.

We now use the framework of section 3 to define three measures of disclosure risk associated with the release of a generic mask, $\tilde{\boldsymbol{f}}$, which correspond to well-known ones proposed on an ad-hoc basis in the literature on statistical confidentiality. In all three examples, $\Phi_A^I = \boldsymbol{F}$ and, since the agency knows the original table, $\pi_A^{(I)}(\phi_A^{(I)} \mid \boldsymbol{f})$ is degenerate at $\boldsymbol{f}$. These examples illustrate how our approach can be used to assess effectiveness of existing criteria. We think of a measure of disclosure (data utility) as sensible if we can obtain it as a result of disclosure scenarios characterized by "natural" choices of users targets, priors, loss functions, etc. At least as important is the application of our framework to define new measures derived from equations (1) and (2) for suitable choices of the input values but this goes beyond the goal of the present paper.

4.1. *Example 1: Disclosure Risk as Tightness of Bounds for Small Cell Counts*

Suppose that the intruder's target is the original table, $\Theta_I = \boldsymbol{F} = (F(1), \ldots, F(m))$, and let the intruder's action space for the problem "estimate $\boldsymbol{F}$" be the m-fold product space (for the purposes of the example we do not require intruder's estimates to lie on the simplex, although in general a rational intruder would include this constraint):

$$\mathcal{N}_I = \overbrace{\mathcal{N} \times \ldots \times \mathcal{N}}^{m\text{-times}}, \;\; \mathcal{N} = \{\, [a, b] \;:\; a \leq b, \quad a, b \text{ non-negative reals}\}.$$

Suppose that when the intruder can define tight bounds for all cells in the table his loss when estimating $\boldsymbol{F}$ is small, whereas if he cannot accurately estimate at least one cell, his loss is large. In particular for a generic $e = (e(1), \ldots, e(m)) \in \mathcal{N}_I$ and $f_j = (f_j(1), \ldots, f_j(m)) \in \mathcal{T}$ assume:

$$L_I(e, f_j) = \begin{cases} \sum_{i=1}^{m} \text{length}\, e(i), & \text{if } f_j(i) \in e(i),\ i = 1, \ldots, m, \\ \infty, & \text{otherwise.} \end{cases}$$

Let $L(i)$ and $U(i)$ be the lower and upper bounds for the ith cell in the original table based on the candidate release $\tilde{\boldsymbol{f}}$, i.e., $L(i) = min\{f_j(i) : f_j \in \mathcal{T}\}$ and $U(i) = max\{f_j(i) : f_j \in \mathcal{T}\}$. For this case, when $\pi_h(\cdot)$ has support $\mathcal{X}$, the intruder's optimal action and uncertainty are, respectively,

$$\hat{\theta}_I = ([L(1), U(1)], \ldots, [L(m), U(m)]), \quad U_I = \sum_{i=1}^{m} U(i) - L(i).$$

Suppose now that the loss that the agency incurs when the intruder takes action a_{rI} takes its minimum when the intruder stops trying to identify the original table ($r = 0$) or when none of the intruder's set estimates of small cell counts in the true table contains the correct value of the cell. Further suppose that the loss increases as the bounds for small cell counts become tighter. This situation corresponds to:

$$L_I^{(A)}(a_{rI}, f_j) = \begin{cases} -\min\limits_{i \in \mathcal{Q}} \text{length}\, \hat{\theta}_I(i), & \text{if } r = 1 \text{ and } \mathcal{Q} \neq \emptyset\,, \\ -n, & \text{otherwise,} \end{cases}$$

where $\hat{\theta}_I(i)$ is the intruder's optimal (set) estimate of $F(i)$ and

$$\mathcal{Q} = \{i \in \{1, \ldots, m\} :\; f_j(i) \in \hat{\theta}_I(i) \text{ and } 0 < f_j(i) < 3\}.$$

If the agency believes that the intruder never stops trying to estimate his target (i.e. $\pi_{T_I}(\cdot)$ is degenerate at nm), then the disclosure risk in (1) becomes:

$$DR(\tilde{\boldsymbol{f}}) = -\min_i \{U(i) - L(i) : 0 < f(i) < 3\}. \tag{3}$$

Choosing this degenerate distribution for the intruder's threshold does not necessarily imply that the agency believes that the intruder always tries to estimate the original table, regardless of his uncertainty, but rather may reflect a conservative attitude based on the worst-case scenario where an intruder always tries to make inference about $\boldsymbol{F}$. The measure in (3) has been discussed on an ad-hoc basis by several authors (e.g., see Dobra, et al., 2002) and is a risk criterion used by many statistical agencies.

4.2. *Example 2: Disclosure Risk as Conditional Probability of the True Table*

Suppose that the intruder's target is the distribution of the original table $\boldsymbol{F}$ and that he uses a logarithmic utility function (Bernardo, 1979):

$$L_I(\hat{P}, f_j) = -log[\hat{P}(f_j)], \quad \hat{P} \in \mathcal{P}, \quad f_j \in \mathcal{X}, \tag{4}$$

where $\mathcal{P}$ denotes the class of all possible distributions with support $\mathcal{X}$. Thus the loss that intruder pays for estimating the distribution of $\boldsymbol{F}$ by $\hat{P}$ when $\boldsymbol{F} = f_j$ (i.e., when the original table is f_j) is a decreasing function of the probability of f_j under $\hat{P}$. Under the loss in (4), the intruder's optimal estimate of the distribution of $\boldsymbol{F}$ is his posterior distribution, and his uncertainty is the entropy of the posterior distribution.

Suppose that the agency pays no loss if the intruder stops trying to estimate the distribution of $\boldsymbol{F}$, and it pays a loss equal to the probability of the true table under the intruder's estimate otherwise,

$$L_I^{(A)}(a_{rI}, f_j) = \begin{cases} 0, & \text{if } r = 0, \\ \hat{\theta}_I(f_j) = \pi_I(f_j \mid \tilde{\boldsymbol{f}}), & \text{if } r = 1. \end{cases}$$

If the agency believes that intruder always tries to estimate the distribution of $\boldsymbol{F}$ no matter what his uncertainty (i.e., if the agency's distribution for the intruder's threshold is degenerate at $log[M(\mathcal{T})]$), then the disclosure risk in (1) is just the (intruder's) posterior probability of the true table $\boldsymbol{f}$ given $\tilde{\boldsymbol{f}}$. If the intruder's prior for $\boldsymbol{F}$ is uniform on $\mathcal{X}$, then from Bayes' Theorem, his posterior given the released table $\tilde{\boldsymbol{f}}$ is uniform on $\mathcal{T}$ and (1) becomes $DR(\tilde{\boldsymbol{f}}) = \pi_I(\boldsymbol{f} \mid \tilde{\boldsymbol{f}}) = 1/M(\mathcal{T})$. Both measures of disclosure have been proposed on an ad-hoc basis in the literature of statistical confidentiality (e.g., see Dobra, 2002). Since they correspond to different assumptions about the intruder's prior for $\boldsymbol{F}$ we can choose between them according to which prior is appropriate for a given problem.

4.3. *Example 3: Disclosure Risk as Fraction of Small Cells Values Correctly Identified*

Suppose that the agency knows that the intruder's target is to identify one cell of the original table, but it does not know which one. If we assume that each cell is equally likely to be the target, we have: $\Theta_I \in \{F(1), \ldots F(m)\}$, and $\Pr(\Theta_I = F(i)) = 1/m$ for $i = 1, \ldots, m$.

Suppose further that the intruder uses a 0-1 loss function, i.e., $L_{Ii}(e, \boldsymbol{f}_j) = 1 - I_{f_j(i)}(e)$. Under these assumptions the intruder's optimal estimate of $F(i)$ is the permissible value $\hat{\theta}_{Ii}$ with highest posterior probability and the intruder's uncertainty is one minus this maximum (posterior) probability.

If the agency's distribution is degenerate at some value t_I^* then, from the agency perspective, the intruder's action is a degenerate random variable that takes values a_{0Ii} or a_{1Ii} depending on whether or not $\Pr(F(i) = \hat{\theta}_{Ii} \mid \tilde{\boldsymbol{f}}) < 1 - t_I^*$. Let δ be a threshold value and let $n_\delta(\boldsymbol{f}_j)$ be the number of cells in $\boldsymbol{f}_j$ such that $\boldsymbol{f}_j(i) < \delta$. Suppose that, when the intruder correctly estimates a small cell value $F(i)$, the agency incurs a loss that is a decreasing function of the number of "small" cell values in the true table and there is no loss if either $F(i)$ is "big" or the intruder's estimate is incorrect. This corresponds to:

$$L_{Ii}^{(A)}(a_{rIi}, \boldsymbol{f}_j) = \begin{cases} m/n_\delta(\boldsymbol{f}_j), & \text{if } r = 1, \hat{\theta}_{Ii} = f_j(i) \text{ and } f_j(i) < \delta, \\ 0, & \text{otherwise.} \end{cases}$$

Then, conditionally on $\Theta_I = F(i)$, the disclosure risk is:

$$DR_i(\tilde{\boldsymbol{f}}) = \begin{cases} m/n_\delta(\boldsymbol{f}), & \text{if } \hat{\theta}_{Ii} = f(i), f(i) < \delta \text{ and } \Pr(F(i) = \hat{\theta}_{Ii} \mid \tilde{\boldsymbol{f}}) > 1 - t_I^*, \\ 0, & \text{otherwise.} \end{cases}$$

and from the mixture version of (1) the (unconditional) disclosure risk is:

$$DR(\tilde{\boldsymbol{f}}) = \sum_{i=1}^{m} \Pr(\Theta_I = F(i)) \cdot DR_i(\tilde{\boldsymbol{f}}) = \frac{\#\,(f(i) < \delta \text{ correctly identified})}{n_\delta(\boldsymbol{f})}. \qquad (5)$$

where in (5) a cell is correctly identified if $\hat{\theta}_{Ii} = f(i)$ and $\Pr(F(i) = \hat{\theta}_{Ii} \,|\, \tilde{\boldsymbol{f}}) > 1 - t_I^*$.

This measure of disclosure has been discussed on an ad-hoc basis by Dobra (2002) and Dobra, et al. (2002). A similar version for microdata is also discussed in Lambert (1993) with $t_I^* = 1$. We next illustrate the implementation of (5) in an example where the released data consists of a set of marginal tables of the original table $\boldsymbol{f}$.

5. UPDATING POSTERIOR DISTRIBUTIONS OVER POSSIBLE TABLES

Now that we have criteria for assessing disclosure risk we can consider the problem posed originally in Section 2, deciding how the agency should respond to requests for specific marginal tables, given the set of tables already released. We do so by looking at the inferences made by the intruder about the possible tables that are consistent with the marginals released to date and we highlight the computational problems that characterize the evaluation of measures of disclosure risk from Section 4.

If the agency has released the l marginals $\mathcal{R} = \{\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_l\}$, and no other information is available about $\boldsymbol{f}$, an intruder knows only that the table $\boldsymbol{f}$ belongs to the set of tables $\mathcal{T}$. [Here $\mathcal{R}$ is equivalent to the masked table $\tilde{\boldsymbol{f}}$ in Section 4.] We treat the population table observation $\boldsymbol{f} = \{f(i)\}_{i \in \mathcal{I}}$ as having been generated from a super-population specified by the random variable $\boldsymbol{F} = \{F(i)\}_{i \in \mathcal{I}}$. Evaluating the disclosure risk associated with releasing $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_l$ by counting the number of tables in $\mathcal{T}$ could create a false sense of security if the probability $\Pr(\boldsymbol{F} = \boldsymbol{f} \,|\, \mathcal{R})$ is high, while $M(\mathcal{T})$ is very large. In this situation, there may be a reasonably substantial probability that the intruder could actually correctly identify the original table $\boldsymbol{f}$. Moreover, we can assess the level of protection for an individual cell count $f(i), i \in \mathcal{I}$, by examining the feasibility interval $[L(i), U(i)]$, where $L(i) = \min\{F(i) : \boldsymbol{F} \in \mathcal{T}\}$, and $U(i) = \max\{F(i) : \boldsymbol{F} \in \mathcal{T}\}$. In many situations we can calculate these bounds directly or using relatively simple algorithms (e.g., see Dobra, 2002, for a general algorithm and Dobra and Fienberg, 2000, 2001 for special cases).

The marginal distribution induced by $\Pr(\boldsymbol{F} = \boldsymbol{f} \,|\, \mathcal{R})$ on the possible values $q \in [L(i), L(i) + 1, \ldots, U(i) - 1, U(i)]$, of a cell $i \in \mathcal{I}$ is given by:

$$\Pr(F(i) = q \,|\, \mathcal{R}) = \sum_{\{\boldsymbol{f} : \boldsymbol{f} \in \mathcal{T}, f(i) = q\}} \Pr(\boldsymbol{F} = \boldsymbol{f} \,|\, \mathcal{R}). \qquad (6)$$

The intruder could infer that the "true" value of cell $i \in \mathcal{I}$ is the value q with the highest conditional probability $\Pr(F(i) = q \,|\, \mathcal{R})$. One could be misled by the fact that the feasibility interval $[L(i), U(i)]$ seems to be wide enough to guarantee the protection of cell count $f(i)$ because the probability of the "true" value $f(i)$ for cell i in (6) might be, in fact, very large and hence $f(i)$ might not be adequately protected.

5.1. *Conditional Distribution of a Table of Counts Under a Log-linear Model*

Suppose the distribution of the cell counts $\boldsymbol{f}$ is multinomial with a fixed total n:

$$\Pr(\boldsymbol{F} = \boldsymbol{f} \,|\, \boldsymbol{\theta}) = \frac{n!}{\prod_{i \in \mathcal{I}} f(i)!} \exp\left[\sum_{i \in \mathcal{I}} f(i) \log \theta(i)\right],$$

where $\theta(i)$ is the probability that an individual cross-classified in table belongs to cell $i \in \mathcal{I}$. The cell probabilities $\boldsymbol{\theta} = \{\theta(i)\}_{i \in \mathcal{I}}$ are constrained to lie within the simplex

$$\Theta = \left\{ \boldsymbol{\theta} : \theta(i) > 0 \text{ for all } i \in \mathcal{I} \text{ and } \sum_{i \in \mathcal{I}} \theta(i) = 1 \right\}. \tag{7}$$

We are more accustomed to working with parameters associated with specific models. We therefore assume that the cell probabilities $\boldsymbol{\theta}$ lie in a space $\Theta_{\mathcal{A}}$ associated with a hierarchical log-linear model $\mathcal{A}$, given by

$$\Theta_{\mathcal{A}} = \Theta \cap \{\boldsymbol{\theta} : \log \boldsymbol{\theta} = A \cdot \boldsymbol{\psi} \text{ for some } \boldsymbol{\psi} = \{\psi(i)\}_{i \in \mathcal{I}} \text{ with } \psi(i) > 0\},$$

where A is the design matrix of $\mathcal{A}$. The introduction of log-linear models for the cell probabilities here is a device and, at the end of this section, we suggest how the results for separate models should be combined.

If $\mathcal{A}$ is the saturated log-linear model, $\Theta_{\mathcal{A}}$ becomes Θ–see equation (7). The conditional distribution of $\boldsymbol{F} = \boldsymbol{f}$ given the released marginals $\mathcal{R}$ under model $\mathcal{A}$ is

$$\Pr(\boldsymbol{F} = \boldsymbol{f} \mid \mathcal{R}, \mathcal{A}) = \int_{\Theta_{\mathcal{A}}} \Pr(\boldsymbol{F} = \boldsymbol{f} \mid \boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta} \mid \mathcal{R}, \mathcal{A})\, d\boldsymbol{\theta}, \tag{8}$$

where $\pi(\boldsymbol{\theta} \mid \mathcal{R}, \mathcal{A})$ is the posterior distribution of cell probabilities given the released margins $\mathcal{R}$ under model $\mathcal{A}$.

Estimating $\Pr(\boldsymbol{F} = \boldsymbol{f} \mid \mathcal{R}, \mathcal{A})$ is difficult because the minimal sufficient statistics of the log-linear model $\mathcal{A}$ might be unknown if we are only provided with the set of marginals $\mathcal{R}$. We need to "augment" the observed data $\mathcal{R}$ to form a complete table $\boldsymbol{F} \in \mathcal{T}$ in order to obtain the minimal sufficient statistics of $\mathcal{A}$. This suggests a data augmentation approach for sampling from the joint density

$$\Pr(\boldsymbol{F}, \boldsymbol{\theta} \mid \mathcal{R}, \mathcal{A}) \propto \Pr(\boldsymbol{F} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \mathcal{R}, \mathcal{A}).$$

Start with $\boldsymbol{\theta}_0 \in \Theta_{\mathcal{A}}$. At the s-th step of the algorithm, do

1. Simulate $\boldsymbol{F}^{(s+1)} \propto \Pr(\boldsymbol{F} \mid \mathcal{R}, \mathcal{A}, \boldsymbol{\theta}^{(s)})$.
2. Simulate $\boldsymbol{\theta}^{(s+1)} \propto \Pr(\boldsymbol{\theta} \mid \mathcal{R}, \mathcal{A}, \boldsymbol{F}^{(s+1)})$.

If we are given the complete table with cell probabilities $\boldsymbol{\theta}^{(s)}$, it no longer makes sense to condition on the log-linear model $\mathcal{A}$. Similarly, given the complete table $\boldsymbol{F}^{(s+1)}$, conditioning on the observed data $\mathcal{R}$ becomes obsolete. Thus

$$\Pr(\boldsymbol{F} \mid \mathcal{R}, \mathcal{A}, \boldsymbol{\theta}^{(s)}) = \Pr(\boldsymbol{F} \mid \mathcal{R}, \boldsymbol{\theta}^{(s)}),$$
$$\Pr(\boldsymbol{\theta} \mid \mathcal{R}, \mathcal{A}, \boldsymbol{F}^{(s+1)}) = \Pr(\boldsymbol{\theta} \mid \mathcal{A}, \boldsymbol{F}^{(s+1)}).$$

We make use of the Markov chain Monte Carlo approach suggested by Diaconis and Sturmfels (1998) for generating draws from the posterior distribution $\Pr(\boldsymbol{F} = \boldsymbol{f} \mid \mathcal{R}, \boldsymbol{\theta}^{(s)})$. This sampling technique relies on the existence of a Markov basis–a finite set of moves or data swaps connecting any two tables with the same marginals.

We need to specify a prior distribution for the cell probabilities that is consistent with the constraints induced by the log-linear model. We take the prior density for $\boldsymbol{\theta}$ to be a constrained Dirichlet prior with hyper-parameters $\boldsymbol{\alpha} = \{\alpha(i)\}_{i \in \mathcal{I}}$ (Schafer, 1997):

$$\pi_{\Theta_{\mathcal{A}}}(\boldsymbol{\theta}) \propto \prod_{i \in \mathcal{I}} \theta(i)^{\alpha(i)-1},$$

for $\boldsymbol{\theta} \in \Theta_{\mathcal{A}}$. It follows that the complete-data posterior density for $\boldsymbol{\theta}$ is

$$\Pr(\boldsymbol{\theta} \,|\, \mathcal{A}, \boldsymbol{F}^{(s+1)}) \propto \prod_{i \in \mathcal{I}} \exp\{[F^{(s+1)}(i) + \alpha(i) - 1] \cdot \log \theta(i)\},$$

for $\boldsymbol{\theta} \in \Theta_{\mathcal{A}}$ and zero otherwise. This is equivalent to the likelihood function for $\boldsymbol{\theta}$ given the table with cell entries $F^{(s+1)}(i) + \alpha(i) - 1$, for $i \in \mathcal{I}$. The constrained Dirichlet prior forms a conjugate class for the multinomial likelihood and hence the posterior of $\boldsymbol{\theta}$ is another constrained Dirichlet prior with hyper-parameters $\boldsymbol{F}^{(s+1)} + \boldsymbol{\alpha}$. We use Bayesian iterative proportional fitting (Gelman, et al., 1995; Schafer, 1997) for simulating random draws from the constrained Dirichlet posterior $\Pr(\boldsymbol{\theta} \,|\, \mathcal{A}, \boldsymbol{F}^{(s+1)})$.

By employing this data augmentation procedure, we can generate a sample $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, ..., $\boldsymbol{\theta}_t$ from the posterior distribution $\pi(\boldsymbol{\theta} \,|\, \mathcal{R}, \mathcal{A})$ and estimate the conditional density of $\boldsymbol{F} = \boldsymbol{f}$ given data $\mathcal{R}$ under model $\mathcal{A}$ from (8) as

$$\Pr(\boldsymbol{F} = \boldsymbol{f} \,|\, \mathcal{R}, \mathcal{A}) \approx \frac{1}{t} \sum_{j=1}^{t} \Pr(\boldsymbol{F} = \boldsymbol{f} \,|\, \mathcal{R}, \boldsymbol{\theta}_j).$$

We are only looking at tables that are consistent with the marginals $\mathcal{R}$; hence we give zero probability to tables that are outside $\mathcal{T}$ by "normalizing" the posterior probabilities in (8) so that they add up to "1":

$$\Pr(\boldsymbol{F} = \boldsymbol{f} \,|\, \mathcal{R}, \mathcal{A}) \longleftarrow \frac{\Pr(\boldsymbol{f} \,|\, \mathcal{R}, \mathcal{A})}{\sum\limits_{\boldsymbol{f}' \in \mathcal{T}} \Pr(\boldsymbol{f}' \,|\, \mathcal{R}, \mathcal{A})}. \tag{9}$$

5.2. *Example*

Table 1 gives a $2 \times 3 \times 3$ table drawn from the 1990 U.S. decennial census public use sample for a local tract, and analyzed previously in Fienberg, Makov, and Steele (1998). Consistent with the discussion in Sections 1 and 4, we act *as if* Table 1 contains population counts. We focus on the four cells containing counts of "1" and "2."

Suppose that the agency releases a pair of 2-way marginals: Race $\times$ Income and Income $\times$ Gender. Table 1 also includes in square brackets the bounds on the cell values resulting from the release of these marginals (Dobra and Fienberg, 2000). Because these marginals are the minimal sufficient statistics of a decomposable log-linear model, there exists a Markov basis that links all $2 \times 3 \times 3$ tables with these marginals (see Dobra, 2002). Table 2 reports the conditional marginal probabilities for the four cells containing small counts induced by conditioning on the saturated log-linear model $\mathcal{A}_1$. We assume a non-informative prior distribution with $\theta(i) = 0.5$ for every cell $i \in \mathcal{I}$. We marked by "–" the values outside the feasibility intervals. By employing the data augmentation algorithm outlined above, we generated a sample of size 500 from $\mathcal{T} \times \Theta_{\mathcal{A}_1}$. The burn-in time for the Markov chain was $1,000,000$. To reduce the correlation between two consecutive draws, we discarded $1,000$ pairs $(\boldsymbol{F}, \boldsymbol{\theta})$ before selecting a new pair in the resulting sample. If the intruder were to "guess" that the true values of the entries for these cells are the values with the highest posterior probability, then his guess would be either incorrect or indecisive for each of the four cells.

Table 1. *Three-way cross-classification of Gender, Race, and Income for a selected U.S. census tract. (Source: Fienberg, Makov, and Steele, 1998). The bounds given in square brackets result from the release of a pair of 2-way marginals: Race × Income and Income × Gender.*

Gender	Race	Income Level $\leq \$10,000$	$> \$10,000$ and $\leq \$25,000$	$> \$25,000$
Male	White	96 [85, 107]	72 [64, 80]	161 [158, 169]
	Black	10 [0, 21]	7 [0, 14]	6 [0, 9]
	Chinese	1 [0, 1]	1 [0, 2]	2 [0, 2]
Female	White	186 [175, 197]	127 [119, 135]	51 [43, 54]
	Black	11 [0, 21]	7 [0, 14]	3 [0, 9]
	Chinese	0 [0, 1]	1 [0, 2]	0 [0, 2]

Tables 2–4. *Marginal conditional probabilities under the log-linear models $\mathcal{A}_1$ (***Table 2***), $\mathcal{A}_2$ (***Table 3***) and $\mathcal{A}_3$ (***Table 4***) for the cells containing small counts in Table 1 induced by releasing the Race × Income and Income × Gender marginals.*

	Table 2			**Table 3**			**Table 4**		
Cell	0	1	2	0	1	2	0	1	2
$(1,3,1)$	0.50	0.50	—	0.64	0.36	—	0.65	0.35	—
$(1,3,2)$	0.43	0.26	0.31	0.44	0.36	0.20	0.39	0.47	0.14
$(2,3,2)$	0.31	0.26	0.43	0.20	0.36	0.44	0.14	0.47	0.39
$(1,3,3)$	0.39	0.25	0.36	0.14	0.34	0.52	0.06	0.37	0.57

5.3. *Updating the Intruder's Posterior Distribution Over Permissible Tables*

In Section 5.1, we calculated the intruder's posterior distribution given a specific log-linear model and noted that we were introducing models as a device. To get rid of this conditioning, we now need to average over the model space, $\mathcal{H} = \{\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_L\}$. The conditional distribution of $\boldsymbol{F} = \boldsymbol{f}$ given the released marginals $\mathcal{R}$ under the family of models $\mathcal{H}$ is (Kass and Raftery, 1995; Madigan and Raftery, 1994):

$$\Pr(\boldsymbol{F} = \boldsymbol{f} \,|\, \mathcal{R}, \mathcal{H}) = \sum_{l=1}^{L} \Pr(\boldsymbol{F} = \boldsymbol{f} \,|\, \mathcal{R}, \mathcal{A}_l) \cdot \Pr(\mathcal{A}_l \,|\, \mathcal{R}), \tag{10}$$

This is an average of the conditional probabilities of $\boldsymbol{F} = \boldsymbol{f}$ under each of the models, weighted by their posterior model probabilities. As we update $\mathcal{R}$ as a result of the release of additional margins, some of the terms in the sum on the r.h.s. of (10) are zero since we need to include only those log-linear models whose minimal sufficient statistics are the same as or include the released margins, $\mathcal{R}$. We note that the posterior probabilities $\Pr(\boldsymbol{F} = \boldsymbol{f} \,|\, \mathcal{R}, \mathcal{A}_l)$ in (10) are not "normalized" as in (9). After calculating $\Pr(\boldsymbol{F} = \boldsymbol{f}' \,|\, \mathcal{R}, \mathcal{H})$, $\boldsymbol{f}' \in \mathcal{T}$, however, we need to "normalize" them to give probability "0" to tables that are inconsistent with $\mathcal{R}$.

The probability of the data $\mathcal{R}$ given the model $\mathcal{A}_l$ is

$$\Pr(\mathcal{R} \,|\, \mathcal{A}_l) = \sum_{\boldsymbol{f}' \in \mathcal{T}} \Pr(\boldsymbol{F} = \boldsymbol{f}' \,|\, \mathcal{R}, \mathcal{A}_l),$$

and thus the posterior probability of model $\mathcal{A}_l$ given data $\mathcal{R}$ is

$$\Pr(\mathcal{A}_l \,|\, \mathcal{R}) = \frac{\Pr(\mathcal{R} \,|\, \mathcal{A}_l) \cdot \Pr(\mathcal{A}_l)}{\sum_{l'=1}^{L} \Pr(\mathcal{R} \,|\, \mathcal{A}_{l'}) \cdot \Pr(\mathcal{A}_{l'})}.$$

5.4. *Example Revisted*

We return to the data in Table 1. There are three log-linear models compatible with the agency release of the two 2-way marginals, Race $\times$ Income and Income $\times$ Gender: (i) the saturated log-linear model $\mathcal{A}_1$, (ii) the log-linear model $\mathcal{A}_2$ of no 2nd order interaction, and (iii) the decomposable log-linear model $\mathcal{A}_3$ for the conditional independence of Race and Gender given Income.

One of the minimal sufficient statistics of $\mathcal{A}_2$, namely Race $\times$ Gender, is not determined by fixing the other two 2-way marginals of Table 1. Table 3 displays the posterior distribution for the possible values of the four cells containing small counts of "1" or "2" given the marginals Race $\times$ Income and Income $\times$ Gender under model $\mathcal{A}_2$. The count of "2" from cell $(1, 3, 3)$ has the largest posterior probability. Hence an intruder using the maximum posterior probability rule would correctly infer one out four values associated with the small counts cells in Table 1.

The minimal sufficient statistics for model $\mathcal{A}_3$ are just the released 2-way marginals. We present the posterior probabilities of the four small count cells under $\mathcal{A}_3$ in Table 4. Here, the intruder would infer the correct value of cells $(1, 3, 2)$, $(2, 3, 2)$ and $(1, 3, 3)$.

The structure of the parameter space clearly makes a difference here since an intruder would not be able to correctly infer with any degree of accuracy the value of any of the four small count cells based on working with the saturated log-linear model $\mathcal{A}_1$. But under the no-2nd-order interaction model, $\mathcal{A}_2$, he could correctly guess one of the four counts and under the conditional independence model, $\mathcal{A}_3$, three of four counts.

Suppose we had assigned these three log-linear models a priori probabilities of 0.22, 0.67, and 0.11, respectively. Combining the results using the model averaging approach of (10), yields posterior probabilities of 0.18, 0.72, and 0.10, respectively. The released marginals (i.e., our data) tend to give more probability to the model $\mathcal{A}_2$ of no-2nd-order interaction–not surprising, since this model fits the data reasonably well whereas the simpler model does not. Table 5 displays the posterior probabilities for the four cells and they are close to those for model $\mathcal{A}_2$. Thus the intruder would not correctly identify the counts in the small cells except for the "2" in the $(1, 3, 3)$ cell.

Table 5. *Posterior conditional probabilities under the family of models $\mathcal{H} = \{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3\}$ for the possible values of the cells containing small counts in Table 1 induced by releasing the Race $\times$ Income and Income $\times$ Gender marginals.*

Cell	0	1	2
$(1, 3, 1)$	0.62	0.38	—
$(1, 3, 2)$	0.44	0.36	0.20
$(2, 3, 2)$	0.20	0.36	0.44
$(1, 3, 3)$	0.15	0.33	0.52

6. SUMMARY AND OPEN PROBLEMS

In this paper we have attempted what we believe to be the first systematic Bayesian treatment of the problem of disclosure limitation for tables of counts, beginning with the trade-off between disclosure risk and data utility, and focusing on intruder efforts to identify small cell counts. The treatment is far from complete, however.

In Section 4, our discussion of data utility was restricted to a single user other than the intruder. But multiple users with differing analytical goals raise new issues. For example, releasing a high-dimensional margin requested by one user might well be "safe," but this action might preclude the release of many other lower-dimensional margins that would be of value to several other users.

Missing from Section 5 is an effort to address the information to the intruder when an agency chooses not to release a requested margin. If the agency is otherwise attempting to maximize the utility of the data for other users, the intruder should understand that the only reason not to release a margin is that when the information in it is combined with the other released margins, the intruder would be able to make "strong" inferences about small cell counts in the full table.

We have also not addressed the alternative strategy to not releasing a requested margin, i.e., perturbing the table (subject of course to the constraints imposed by the already released margins) and releasing the margin from the perturbed table. This is a form of data transformation or mask in the spirit of Section 3. Clearly to do such perturbation in an efficient manner, the agency would do well to compute its posterior distribution over the parameters of the super-population space, and then draw tables from that distribution. This would be akin to the approach suggested in Fienberg, Makov, and Steele (1998) or the multiple imputation approach of Raghunathan and Rubin (2001). But then the intruder needs to update his distribution over the space of possible tables in a somewhat different fashion than that in Section 5.

As we noted in Section 1, small counts in a sample table may not necessarily correspond to small counts in a population table. Thus we need to adapt the strategies for assessing disclosure risk from Section 5 to deal with sample tables. Intuitively, as the sampling fraction gets smaller we expect disclosure risk to go down. But this may not be sufficient protection.

The U.S. decennial census files from which that 3-way table was extracted contain 53 categorical variables, cross-classified at multiple levels of geography. The kinds of computations we were able to implement on the 3-way table in Section 5 do not necessarily scale well for such large class-classifications. Many of the calculations in Section 5 have a remarkable similarity to those involved in Bayesian model search with hierarchical log-linear models and especially the subclasses of decomposable and graphical models, just as the work on bounds for contingency tables in Dobra and Fienberg (2000, 2001) had intimate links to decomposable and graphical models. Thus tools such as those associated with the hyper-Markov laws in Dawid and Lauritzen (1993), and the suite of expert system tools in Cowell, et al. (1999) may prove useful as we develop scalable approaches to disclosure limitation in large tables of counts.

ACKNOWLEDGMENT

This research has been supported in part by National Science Foundation Grant No. EIA-9876619 to the National Institute of Statistical Sciences, and by a Marie Curie Fellowship of the European Community program "Improving The Human Research Potential" under the contract number HPMFCT-2000-00463 to the Universitat de València. We thank our colleagues on these projects, Susie Bayarri, George Duncan, Alan Karr, Steve Roehrig, and Ashish Sanil, for helpful ideas and comments.

REFERENCES

Bernardo, J.M. (1979). Expected information as expected utility, *Ann. Statist.* **7**, 686–690.

Chen, G. and Keller-McNulty, S. (1998). Estimation of identification disclosure risk in microdata, *J. Official Statist.* **14**, 79–95.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems*. Berlin: Springer.

Dawid, A. P. and Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models, *Ann. Statist.* **21**, 1272–1317.

DeGroot, M. H. (1962). Uncertainty, information and sequential experiments, *Ann. Math. Statist.* **33**, 404–419.

Diaconis, P. and Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *Ann. Statist.* **26**, 363–397.

Dobra, A. (2002). Statistical Tools for Disclosure Limitation in Multi-way Contingency Tables. Ph.D. Thesis, Department of Statistics, Carnegie Mellon University.

Dobra, A. and Fienberg, S. E. (2000). Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *Proc. of the National Academy of Sciences* **97**, 11885–11892.

Dobra, A. and Fienberg, S. E. (2001). Bounds for cell entries in contingency tables induced by fixed marginal totals. *Statistical Journal of the United Nations ECE* **18**, 363–371.

Dobra, A., Fienberg, S. E., Karr, A. and Sanil, A. (2002). Software systems for tabular data releases, *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*,(to appear).

Dobra, A., Tebaldi, C. and West, M. (2002). Reconstruction of contingency tables with missing data. Manuscript.

Doyle, P., Lane, J. Theeuwes, J., and Zayatz, L. eds.) (2001). *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, Amsterdam: Elsevier.

Duncan, G.T., Fienberg, S.E., Krishnan, R.,Padman, R. and Roehrig, S.F. (2001). Disclosure limitation methods and information loss for tabular data. In *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, (P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz, eds.), Amsterdam: Elsevier, 135–166.

Duncan, G. T., Keller-McNulty, S., and Stokes, S. L. (2001). Disclosure risk vs. data utility: The R-U confidentiality map, *Tech. Rep.*, Los Alamos National Laboratory, NM..

Duncan, G. T. and Lambert, D. (1986). Disclosure-limited data dissemination (with discussion), *J. Amer. Statist. Assoc.* **81**, 10–28.

Duncan, G. T. and Lambert, D. (1989). The risk of disclosure for microdata, *J. Business and Economic Statist.*, **7**, 207–217.

Duncan, G. T., and Pearson, R. B. (1991). Enhancing access to microdata while protecting confidentiality: Prospects for the future (with discussion), *Statist. Sci.* **6**, 219–239.

Fienberg, S. E. (1994). Conflicts between the needs for access to statistical information and demands for confidentiality. *J. Official Statist.* **10**, 115–132.

Fienberg, S. E. and Makov, U. E. (1998). Confidentiality, uniqueness, and disclosure limitation for categorical data. *J. Official Statist.* **14**, 385–397.

Fienberg, S. E. and Makov, U. E. (2001). Uniqueness and disclosure risk: Urn models and simulation. *Research in Official Statistics* **4**, 23–40.

Fienberg, S. E., Makov, U. E., and Sanil, A. P. (1997). A Bayesian approach to data disclosure: Optimal intruder behavior for continuous data. *J. Official Statist.***13**, 75–89.

Fienberg, S. E. and Makov, U. E., and Steele, R. J. (1998). Disclosure limitation using perturbation and related methods for categorical data (with discussion). *J. Official Statist.* **14**, 485–511.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*. London: Chapman and Hall.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors, *J. Amer. Statist. Assoc.* **90**, 773–795.

Lambert, D. (1993). Measures of disclosure risk and harm, *J. Official Statist.* **9**, 313–334.

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window, *J. Amer. Statist. Assoc.* **89**, 1535–1546.

Raghunathan, T. E. and Rubin, D. B. (2001). Multiple Imputation for statistical disclosure limitation. *Tech. Rep.*, ..

Rubin, D. B. (1993). Satisfying confidentiality constraints through the use of synthetic multiply imputed microdata. *J. Official Statist.* **9**, 461–468.

Samuels, S. M. (1998). A Bayesian, species-sampling-inspired approach to the uniqueness problem in microdata disclosure risk assessment. *J. Official Statist.* **14**, 373–383.

Skinner, C. J. and Holmes, D. J. (1998). Estimating the re-identification risk per record in microdata. *J. Official Statist.* **14**, 361–372.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.

Trottini, M. (2001). A decision-theoretic approach to data disclosure problems. *Research in Official Statistics* **4**, 7—22.

Trottini, M. and Fienberg, S. E. (2002). Modelling user uncertainty for disclosure risk and data utility. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, (to appear).

Willenborg, L. and De Waal, T. (1996). *Statistical Disclosure Control in Practice.* Berlin: Springer.

Willenborg, L. and De Waal, T. (2001). *Elements of Statistical Disclosure Control.* Berlin: Springer.

DISCUSSION

CHRISTOPHER MEEK *(Microsoft Research, USA)*

Congratulations to the authors for an interesting and thought-provoking paper. Methods for managing and assessing disclosure risks for information is an area of growing importance due to increasing amount of confidential information available electronically. It is useful to note that while the authors have used examples from census surveys for illustration, they could have chosen from a variety of alternative scenarios including disclosure of information from records and surveys on medical procedures, product purchases, internet usage, and subscriptions.

The paper provides a general decision-theoretic framework for assessing and managing disclosure risk. In addition, the authors describe techniques that begin to address the problem of inference about a joint table given released sub-tables; a problem that is central to using their framework. The goal of the general decision-theoretic framework is to allow an *agency* to balance the risks and benefits of disclosing information. The authors capture the benefit and risk with data-utility and data-risk functions, respectively, and evaluate the alternative potential releases by combining the data-utility and data-risk functions. The authors suggest combining these function with a thresholded decision rule. This suggestion seems odd; whereas this rule does utilize both the data utility and the data risk it doesn't balance them. If two sub-tables have the same data utility why not choose the one with the lower data risk?

It is most natural to consider the problem faced by such an agency as a multi-step decision problem or game-theoretic problem but, to simplify, the authors consider a myopic one-shot decision problem: should the agency release a sub-table $\mathcal{T}$ given previously released tables $\mathcal{R}$. The basic tactic employed by the authors in defining the benefits and risks is to equate the benefit (data utility) with the negative expected loss of the scientist and the data risk with the expected loss of the intruder. In their analysis, they ignore the game-theoretic aspects of the loss function; in actuality we would expect that the actions of the *intruders* and *scientists* are informed by their understanding of the system and their beliefs about the other participants, future requests and so on. The authors rather consider simplified loss functions that allow one to decompose the larger problem into more tractable sub-problems. Unfortunately it is unclear whether this approach will allow one to create reasonable data utility and data risk functions. A fundamental question that the authors do not discuss is how one should go about improving the data utility and data risk functions. Should one interview intruders and scientists? Perhaps more importantly, how does one combine the loss functions of multiple intruders and scientists? One cannot simply add loss functions.

The authors demonstrate that they can use their framework to obtain data risk functions that have been used in the literature and thus ascribe a loss function to the putative intruder. This

work demonstrates how extreme and unreasonable loss functions must be to reproduce the data risk functions, and points to the need to identify more realistic loss functions.

The development of more realistic loss functions in combination with the framework has other potentially useful applications. For instance, there are a number of alternative actions that the agency might take to protect data confidentiality when given a request for data. The authors focus on the case in which the agency simply must decide which table, if any, to release. Other methods for releasing data such as alteration, masking, or releasing ranges of values could be compared and evaluated using the framework when coupled with realistic loss functions.

Throughout the paper the authors equate small counts with high data risk. At an intuitive level this seems reasonable because counts of one or two might be identifiable and lead to the release of identifiable confidential information either directly through information released by the agency or indirectly through the linking of information release by the agency and other information sources. The following examples illustrate potential problems of equating small counts with high data risk.

First, consider a sub-table with attributes hair color, height and town. Let us suppose that there is one small town in which there is one person who has red hair and who is 7 feet tall. Clearly this person is identifiable (just go to the town and you can probably find the person!), however, this information alone is not confidential.

Second, consider a sub-table with attributes corresponding to a person's favorite movie, favorite book, and favorite color. In this situation, there are certainly going to be some entries that will have small counts. Unlike the previous case, these attributes do not lead to identifiability because we do not wear our preferences on our shirt sleeves.

Finally, consider the sub-table given in Table 1 of Income (I) and Hair Color (H) for town X. In this table, none of the entries is small in the sense discussed by the authors. Nonetheless, a release of such a table does seem to provide a release of information—that is, all blondes have high income in town X—and the blondes in town X are identifiable (again, just go to the town). If income is deemed to be confidential, then the release of this sub-table would be a release of identifiable confidential information. This example illustrates that the release of confidential information is not limited to the case of small counts.

Table 1. *Example of disclosure without small counts.*

	Town = X	
	I = High	I = Low
Black	4	4
Red	4	4
Blond	4	0

The last section of the paper is devoted to a specific technical problem central to the application of the general framework: the problem of computing a posterior over the joint table given a set of released sub-tables $\mathcal{R}$. This problem is not novel to the analysis of disclosure risk; it also arises in sociology, political science, spatial epidemiology and ecology where it is know as the ecological inference problem. The authors choose to attack the problem by considering the joint table to be a sample from an infinite super-population. In the simple case, the authors then consider a fixed log-linear model and utilize a two-step Gibbs sampler with one step requiring MCMC and the other a "Bayesian iterative proportional fitting" procedure. In the more complicated case, the author combines the Gibbs sampling approach with model averaging over a set of alternative log linear models.

For the MCMC step, the authors are to be congratulated on what is probably the first Bayesian use of the Markov basis results of Diaconis and Sturmfels (1998). They construct a Markov chain to sample tables from the posterior of tables that satisfy the released sub-tables. Unfortunately, the computation of the Markov basis is very expensive, and precious little is known about mixing times of Markov chains constructed from them. An alternative approach to sampling a table from the posterior of tables that satisfy the released sub-tables is based on sequential cell sampling (Dobra, Tebaldi and West 2002). In this approach, one sequentially samples every cell and updates the upper and lower bounds. Clearly, this method can also be extremely expensive computationally (potentially exponential in cost). Given these computational difficulties, which of these methods do you think will be able to scale to large problems? Do we need to develop some new class of approaches?

The specific approach advocated by the authors is not completely specified. When performing model averaging, neither how one should go about selecting the set of alternative models, nor how one ought to compute the model posteriors for these alternatives is clear. What is the precise recipe? For some classes of models, for instance, non-decomposable and non-graphical models, the computation of the marginal likelihood is non-trivial even in the case of complete data. The situation in which one is computing the marginal likelihood of released sub-tables $\Pr(\mathcal{R} \mid \mathcal{A}_l)$ is certainly more challenging. If one wishes to include such alternative models, it may be useful to combine the model averaging with the Gibbs sampling by using a reversible jump MCMC approach.

In addition, given the focus on the small counts, it would seem sensible to carefully consider the sensitivity of the results.

Finally, the authors consider the problem of computing a posterior on the joint table given a set of released sub-tables when, in fact, there is more information to consider. Perhaps the most important information that is not accounted for is the refusal to release. Consider an agency that uses the rule that it will release no sub-table that makes a cell unique. In this case, if a request for a two-by-two table with a grand total of four is refused after both of the margins have been released, then there is only one table that is possible. Because the rule used by the agency is a function of the actual joint table, the tactic used by the authors of using a super-population to make inferences about the joint table complicates the inferential process of conditioning on the information provided by the refusal to release a sub-table. Whereas assumptions such as the super-population assumption allows one to use many standard statistical tools, perhaps it is worthwhile considering alternative approaches to this problem.

REPLY TO THE DISCUSSION

We thank Chris Meek for insightful comments and questions and for his clear appreciation of the complexity of the problem we are attempting to address. As he notes, issues of confidentiality and disclosure limitation arise in many different contexts. Most recently we have begun to consider them as part of a research effort in computer security. We organize our response according to the three components of the paper: understanding disclosure risk for categorical data, aspects of our decision-theoretic approach, and the computing the posterior distribution over feasible tables given data releases.

Understanding Disclosure Risk. Meek asks why we equate small counts with high data risk. As we noted in our introduction, we distinguish, as does the confidentiality literature, between *identity* disclosure (through uniqueness in the population) and *attribute* disclosure (see also Duncan, et al., 1993). His example of identity disclosure involves a small town with only one person with red hair and who is seven feet tall and he notes that this information alone may well not be confidential. This is true as far as he goes. But if the two variables are only a subset

of those involved in our multi-way table, then all of the values on the other variables in the data base on this individual are disclosable and this may indeed be a serious problem. It is largely for this reason that statistical agencies and others who gather data promise those that supply it that the data will not be released in a form that allows individuals to be identified. Protecting identities then becomes by definition a primary confidentiality issue.

Meek's second example, of a small town where all blonds have high income, is an example of attribute disclosure–we don't learn the identity of any specific blond but we learn that they all have high income. We explicitly set such an issue aside in our paper, although it turns out that we could easily adapt the discussion of both decision-theoretic criteria and intruder identification to deal with this situation. For example, by looking at bounds on all cells and then focusing on small cells for identity disclosure and on cells whose values almost equal one or more of the released marginal values to which it sums for attribute disclosure.

Aspects of Our Decision-theoretic Approach. Meek asks about our use of the threshold rule which is widely used in the current literature on statistical confidentiality. Our simplified version implicitly assumes that for a fixed threshold α there is a unique data mask that maximizes data utility among all the available masks with disclosure risk below α (this is the most common case in applications). Clearly this simplification does not perform well when there are two or more masks that maximize data utility (the example raised by Meek). We implicitly assume in this case, however, that the mask with minimum risk should be selected. Thus, the rule actually balances disclosure risk and data utility. A general threshold rule, that solves the incoherence highlighted by Meek, could be stated as follows: *Let $\mathcal{D}$ be the class of all available masks and let $\mathcal{D}_\alpha$ be the subclass of $\mathcal{D}$ containing all masks with disclosure risk below α. Finally, let $\mathcal{D}_\alpha^{OPT}$ be the class of masks in $\mathcal{D}_\alpha$ that maximize data utility. If $\mathcal{D}_\alpha^{OPT}$ is empty, release no data. Otherwise release the mask in $\mathcal{D}_\alpha^{OPT}$ with minimum disclosure risk.*

The suggestion that we formalize the disclosure limitation problem as a multi-step decision problem or game-theoretic problem is appealing at a first glance and it seemed to us the most natural choice when we first approached the problem. In the end, we preferred the one-shot decision problem formalization because: (i) the real-world disclosure limitation problem is actually a one shot problem: the agency release the data, the users make their move (estimate of their targets); (ii) there is a strong asymmetry of information between the agency and the users. The agency knows the original data, the alternative masks available, the mask that has been used, the users' targets and loss functions, and it fixes the threshold value for the maximum tolerable disclosure risk. On the other hand, users only know that the released data have been obtained using a particular masking technique and sometimes they ignore the same threshold rule. Users understanding of the system is quite limited and their beliefs about other participants (the agency in particular) are so vague that they are of little inferential value. Thus, the one-shot decision approach is not a myopic simplification of the problem but rather a very good approximation of how things works in reality. The key point is that in order to fill its mission of data dissemination, the agency does not need to reveal its strategy to the users but simply provide the data in a form that is useful for the *scientist* and safe from the attack of an *intruder*.

Meek questions how we should approach defining suitable loss functions for the intruder and the researcher. Many statistical agencies either ignore the uses of the data they release or claim that taking into account all users' targets is impossible. Instead they use heuristic ad-hoc measures that try to preserve basic features of the original data while preserving confidentiality. Our framework clearly shows that defining measures in such a way does not avoid the problem of defining targets and loss functions. Rather ad-hoc measures turn out to correspond to very specific (and often unrealistic) choices of them. While defining users' targets and loss functions is not easy, it is necessary and unavoidable if we want suitable measures of disclosure risk and

data utility. The idea of interviewing the users is appealing but difficult to implement for the intruders that matter the most. For scientists the job is easier, however, since, most statistical agencies already keep track of requests from users and design forms of data release that can meet these requests. Thus, if an agency knows that that the principal use of the data is the estimation a set of statistical models, then it can design suitable releases to preserve the statistical features of the parameters of these models. To define intruders' targets and loss functions the agency needs to reflect on what information in the original data needs to be protect and then consider a range of possible attacks an intruder might attempt to disclose this information.

Combining the loss functions of multiple users is also relevant. For disclosure risk, a reasonable solution is to use a worse case strategy and adopt the maximum disclosure risk measured for different targets on a comparable scale. The data utility case is more complex. We could consider linear combinations of users' loss, with different weights for different users and we are currently working on this problem. The current version of the framework nonetheless allows us to identify situations where there exists a suitable data mask that can meet the needs of several users of the data while preserving confidentiality and situations where the targets of the multiple scientists are not compatible (i.e., masks suitable for one user are almost useless for another).

Computing the Posterior Distribution over Feasible Tables. As Meek notes this problem also arises in other contexts such as ecological inference (e.g., see King, 1997) as well as tomography image reconstruction (e.g., see Gelman, 1989), although in these other settings we are typically interested in some form of model whereas in the disclosure case we are not necessarily.

Meek asks which of the methods we explore holds the promise to scale to large problems and whether we should consider developing some new classes of approaches. The approach proposed by Diaconis and Sturmfels is not likely to scale because of the huge complexity of the computations required to compute a Markov basis. Although there exists a formula for dynamically generating a Markov basis in the decomposable case (Dobra, 2001) and ways to significantly speed up the computations in some other special cases (see, for example, the divide-and-conquer algorithm of Dobra and Sullivant (2002)), there is little chance that a Markov basis associated with a higher dimensional table and an arbitrary set of marginals will be easily computable in the near future. The sequential cell sampling approach of Dobra, Tebaldi and West (2002) offers a possible answer to this general problem, but this method relies on the computation of bounds, which is itself a hard problem for large tables. Therefore neither approach seems to scale to large tables and we need to explore new avenues as well as approximations that use reasonable amounts of computation.

Meek argues that we did not completely specify our approach in computing the posterior distribution over the space of feasible tables. Our example was small enough that, at least for it, we believe that there was no ambiguity, and we tried to make clear how one can compute or at least approximate the posterior distribution for each model. Nonetheless, there is clearly a major issue here for high-dimensional tables. Ideally we need to average over the space of all possible hierarchical log-linear models "consistent" with the released marginals, but, as Meek notes, the computations appear daunting without the added complexity of simultaneously using a reversible-jump MCMC approach. We are currently exploring approximations based on decomposable models but we also need to be able to compute the posterior distribution for the model whose minimal sufficient statistics correspond to the released margins, so we do not as yet have any recipe.

We agree with Meek's suggestion that it would be sensible to consider the sensitivity of our results to the specification of the model priors. We agree, and expect to explore alternatives

such as the compatible priors associated with hyper-Markov laws.

Finally, we recognize the need to take into account the information supplied to an intruder through the refusal to release tables. Meek presents an extreme example, and the more likely scenario is that an agency is willing to release most lower-order margins and its refusals will occur as the dimensionality grows. Our bounds work (e.g., see Dobra, et al. 2002) is clearly consistent with this scenario, as is the U.S. Census Bureau's American FactFinder system which has been constructed to release at most 3-way tables. But this is one of the reasons why the option of perturbing a table before releasing a new "masked" margin is an option worthy of exploration.

ADDITIONAL REFERENCES IN THE DISCUSSION

Dobra, A. (2001). Markov bases for decomposable graphical models. *Tech. Rep.*, National Institute of Statistical Sciences.

Dobra, A. and Sullivant, S. (2002). A divide-and-conquer approach for generating Markov bases of multi-way tables. *Tech. Rep.*, National Institute of Statistical Sciences.

Duncan, G. T., Jabine, T. B., and Wolf, V. A. de (Eds.). (1993). *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*. Washington, DC: National Academy Press.

Gelman, A. (1989). Constrained maximum entropy methods in an image reconstruction problem, in *Maximum Entropy and Bayesian Methods*, (J. Skilling, ed.), Dordrecht: Kluwer, 429–435.

King, G. (1997). *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton: University Press.