\documentclass[10pt]{article}
\input{Unix Fonts}
input{ISBA}
\begin{document}

\paper\pin{1}

\bayes{}{}

\tit {Towards a Bayesian Perspective on Statistical Disclosure Limitation}

\aut{LAWRENCE H. COX}

\loc{U.S. Environmental Protection Agency, USA}

\bas

National statistical offices and other organizations collect data on individual subjects (persons, businesses, organizations), typically while assuring the subject that data pertaining to them will be held confidential. These data provide the raw material for statistical data products (tabular summaries, microdata files comprising data records pertaining to individual subjects, and, potentially, public statistical data bases and statistical query systems) which the statistical office disseminates to multiple, broad user communities. Statistical disclosure limitation (SDL) refers to the problem and methods for thwarting re-identification of a subject and divulging the subject's confidential data through analysis or manipulation of disseminated data products. SDL methods abbreviate or modify the data product sufficiently to thwart disclosure. SDL problems are typically computationally demanding; several have been shown to be NP-hard. Many SDL methods draw upon statistical, mathematical or optimization theory, but at the same tim heuristic and partial approaches abound. Contributions from a Bayesian perspective have been few but are increasing. A strong theoretical connection between definitions of statistical disclosure, measurement of disclosure risk, and evaluation of SDL methods is lacking. This suggests opportunities for Bayesian and hierarchical approaches. Selected opportunities and associated SDL methodological issues are discussed.

\eas

\key{disclosure risk; synthetic data; multiple imputation; Bayesian model averaging; E-M algorithm; exact linear bounds; MCMC computation}

\section {1.}{Introduction}

Bayesian and hierarchical methods have been used with success in several areas of official statistics and public policy. These methods are particularly well-suited to assessment of condition and risk, e.g., environmental and ecological modeling, health risk assessment, and assessment of computer models, often in combination with statistical methods for spatial, temporal, spatio-temporal and trends analysis.

\it{Statistical disclosure limitation} (SDL) refers to a suite of methods designed to thwart efforts by unauthorized third parties to identify subjects of statistical inquiries and infer confidential data pertaining to them. Most SDL methods modify or abbreviate the data product to the point where subject data has been sufficiently \it{masked}. SDL methods used by national statistical offices (NSOs) include: for frequency count tabulations, rounding or introducing random noise into counts; for tabulations of magnitude data, e.g., monthly sales, cells that represent unacceptable risk of disclosure are suppressed from publication, together with sufficiently many it\{complementary suppressions} to ensure that original suppressions cannot be reconstructed or narrowly estimated; and, for statistical microdata, viz., data files whose records correspond to individual responding units, a variety of local suppression, recategorization, rounding, perturbing and ad hoc methods are used. In addition, replacing original data by model-generated data has been suggested. Federal Committee on Statistical Methodology (1994) for provides discussion of confidentiality issues in U.S. official statistics and a synopsis of SDL methods. Only limited use of Bayesian and hierarchical methods has been made in statistical disclosure limitation, with a few notable exceptions.

I was invited to ISBA 2000 to speak on my perspective on opportunities for use of Bayesian methods in statistical disclosure limitation. I chose to include hierarchical and likelihood methods with Bayesian approaches. From the outset, it should be clear that I am a consumer, not a producer, of Bayesian methods.

The term "disclosure limitation" has emerged from earlier terminology including "disclosure protection" and "disclosure control" in recognition of the fact that the release of statistical data inevitably implies something about the characteristics of the subjects from which the data are derived. This terminology implies to me the existence of a quantitative measure (most likely, a continuous measure) of disclosure on which the degree of "limitation" can be assessed. Unfortunately, this intuitively valid notion has failed to find its way into rigorous implementation in most aspects of disclosure limitation. This lacuna in the research fabric is the main motivation for my comments.

The next two sections highlight selected areas of statistical disclosure limitation methodology where introduction of Bayesian and hierarchical methods appears to me to be promising. Section 2 deals with applications already ongoing, and focuses on opportunities related to increased or enhanced use of these methods. Section 3 deals with areas where Bayesian and hierarchical methods have not been applied, and focuses on promising avenues of research. A research direction in MCMC computation, combining Bayesian statistics and mathematical optimization, is proposed. Section 4 contains concluding comments.

\section{2.}{Expanded Use of Existing Methods}

In Sec. 2.1 we present examples of current and past Bayesian and hierarchical approaches to problems in statistical disclosure limitation. These examples were selected based on opportunities we perceive for increased or extended work along similar lines, as discussed in Sec. 2.2. No attempt is made to summarize these approaches fully nor to provide an exhaustive list of Bayesian and hierarchical approaches to SDL.

\subsection {2.1} {Existing Methods}

Following work in the 1970s by O. Frank and C.-M. Cassel on distribution-based attacks on exact and approximate disclosure in tabulations, the introduction of Bayesian and hierarchical approaches to statistical disclosure limitation began in the 1980s. Its continuation at ISBA 2000 included two invited session on SDL involving Bayesian data aggregation (Kokolakis and Nanopoulos), Bayesian multiple imputation (Raghunathan and Rubin), comparing masked and synthetic microdata (Duncan and Keller-McNulty), modeling population uniques (Fienberg and Makov), and hierarchical Bayesian models for producing synthetic microdata from economic surveys (Franconi and Stander).

Duncan and Lambert (1986, 1989) developed Bayesian risk models for tabulations and microdata files. The authors demonstrate a correspondence between plausible uncertainty functions and commonly used operational rules to define statistical disclosure, thereby providing a foundational link between rules used in practice and quantifiable notions of disclosure risk (thus, the first step towards an operational realization of our intuitive notion). Tools of this sort enable the development of disclosure rules based on \it{nominal disclosure}, viz., how close, based on prior information, the attacker can come to confidential information, or on \it{relative disclosure}, viz., how much the attacker's prior knowledge of confidential information increases following release of the data product.

Paass (1985) employed statistical matching methods to simulate disclosure attack on a public use microdata file using a matching file. Paass simulated errors in matching variables from standard distributions. While later research pointed to the conclusion that actual inconsistencies between files do not obey simple distributional paradigms, and that real files are actually harder to match than simulated ones, Paass' work did demonstrated that massive but nevertheless simple random perturbation of data values is unlikely to provide meaningful disclosure limitation.

Microdata files are typically created by sampling larger (in some cases, population) files. The degree and complexity of the sampling provides some disclosure limitation. At greatest risk are sampled subjects whose records are unique in the sample file. Skinner and Holmes (1992), Fienberg and Makov (1998), and Samuels (1999) are concerned with the consequent problem of inferring that a sample-unique is in fact a population-unique. Skinner and Holmes discuss but do not employ a fully Bayesian modeling framework; the other authors do. Further work reported at ISBA 2000 by Makov aims to improve Samuel's approach.

Rubin (1993) and, at ISBA 2000, Raghunathan and Rubin, believe that acceptable levels of disclosure risk for microdata files are not achievable using standard SDL methods, and propose releasing only \it{synthetic microdata}. The authors propose a methodology based on constructing a statistical model from the original microdata and releasing versions of the file containing model-generated imputations in place of original data values. Multiply imputed versions of the file are released so that users can estimate the increase in variance caused by imputation.

Fuller (1993) assess the vulnerability of masked microdata to attack. Little (1993) investigates methods for analyzing disclosure-limited data. Lambert (1993) distinguishes between \it{disclosure risk}, viz., probability of identifying a subject from the data, and \it{disclosure harm}, viz., likelihood of attributing confidential information or sensitive characteristics to the subject.

\subsection {2.2}{Expanded and Related Methods}

I see three areas where work on Bayesian methods applied to statistical disclosure limitation could be extended.

Bayesian methods could be used to develop formal models and definitions of disclosure risk. A disclosure rule assumes something about prior information available to the attacker, even if this is not stated explicitly. For example, an individual's income or a business's receipts can be estimated with confidence to within some percentage, even if this percentage is large. Some of the characteristics of an individual, e.g., gender, residence, profession, are likely to be publicly known; others, e.g., age, can be approximated; and others, e.g., nationality, marital status, can be inferred according to some probability distribution. Quantitative disclosure limitation criteria, such as it\{minimum population threshold} (the population-based size of the smallest geographic area identified) or SDL methods, such as rounding, can be related to probabilistic statements. Expected values can be computed for population-level cross-classified tabulations. All of this, plus public information can be used to develop reasonable prior distributions on attacker knowledge. Within a Bayesian framework, posterior predictive distributions can be estimated for SDL-related statistics such as expected number of population uniques or the percentage within which a competitor can estimate business data. This in turn enables assessment of the actual disclosure protection provided by the original disclosure rule and disclosure limitation criteria and methods. Finally, with this experience and such models, the NSO can evolve models and improved definitions of disclosure and disclosure risk that relate prior and specialized information to posterior inference. The framework of Lambert (1993) should be explored in this context.

More work on classifying population uniques and small domains would be beneficial. NSOs would benefit greatly from experience along these lines that could help quantify conditions under which population uniqueness or small domains can be inferred with confidence. Examples of potential conditions include the number of cross-classifying variables permitted or the distribution of marginal totals. As just indicated, it would also be useful to expand the focus of investigation from uniqueness to small domains. In addition to salient individuals, it is important to quantify and compare disclosure risk for a randomly selected individual and that for the average individual, as the approach to and degree of disclosure limitation is likely to differ between these cases.

Rubin's arguments for releasing synthetic data are compelling, and deserve further investigation. A principal criticism of model-generated data is that it is dependent on the completeness and representativeness of the model, e.g., interactions not modeled cannot be examined using synthetic data. Perhaps this can be addressed by use of \it{Bayesian model averaging} so that an entire class of models can be simulated. An important question is then whether NSOs should simply release the model(s). Work is needed on the practical analytical use of synthetic data, and on the kinds of ancillary information the NSO could provide to support and to fill gaps not covered by the model, e.g., \it{contextual variables}. Discussions to date of synthetic data have focused on microdata. For tabular data, the release of interval data or synthetic tabulations in lieu of data with suppressions should be investigated. Finally, the predominant issue of quantifying and evaluating the effects of disclosure limitation on data use it seems to me fits neatly within a hierarchical Bayesian framework, whether those data are disclosure-limited or synthetic.

Each of the above three proposals deals with cutting-edge research. To be fully useful and evaluated by NSOs, the results of such research, viz., models and methods, should be made available in well-documented software. In addition to helping familiarize NSO personnel with the methodology and its limitations, software enables transfer of the technology from developer to user. This is of particular importance because realistic testing of SDL methods requires actual, not simulated data, often unavailable to researchers outside the NSO.

\section {3.} {Opportunities for Bayesian and Hierarchical Approaches in Statistical Disclosure Limitation}

This section presents selected problems in SDL and associates to each Bayesian and hierarchical approaches that in my view promise to offer improvement towards solving the problem. Section 3.1 deals with a range of problems of a general sort. Section 3.2 deals with an important problem for tabular data.

\section{3.1} {Selected Opportunities}

Paass (1985) provided the first serious attacker simulation in the literature. Unfortunately, relatively few studies of this sort have since been performed, and we list it as new, rather than expanded, work. Bayesian approaches are natural here: attacker scenarios can be expressed as prior distributions, disclosure limitation criteria and methods can be applied to the original data or represented as constraints, and posterior predictive distributions can be estimated and characterizations of disclosure risk quantified. For microdata, prior knowledge, ancillary information such as published tabulations, and record linkage techniques could be simulated. For tabulations, linear structural equations would be incorporated. For statistical data base query systems, these features as well as resampling could be simulated and examined. From these simulations would arise an in depth, realistic understanding of the relationship between disclosure rules, disclosure limitation strategies and actual disclosure risk.

It would be easy to incorporate measurement and sampling error within a Bayesian model of a data product and its disclosure limitation. This would enable assessment of the degree of disclosure limitation provided by factors already present in original data.

Multi-dimensional contingency tables are a standard data product, and are staples of a statistical data base. For confidentiality reasons, internal entries and some marginal totals may be partially or completely suppressed, rounded or perturbed prior to release. \it{Iterative proportional fitting}, a statistical algorithm based on likelihood methods, can be used to impute missing or distorted values, subject to known marginals. More general types of missing data can be estimated via the E-M algorithm. Limited experience shows that such approaches often come close to original data. It is important to investigate further the degree of this similarity and the reasons why.

Bayesian models can be used to simulate important confidentiality scenarios. The usefulness of disclosure-limited microdata files can be enhance by providing ancillary information such as contextual variables. The effects on confidentiality of various forms of ancillary information, or of record linkage between two files, needs to be simulated. Statistical data base query systems, discussed in the literature during the 1990s, are sure to emerge this decade. A potential problem in this environment is it\{gridlock}, viz., reaching a point where further release of data would lead to disclosure. Bayesian models could be used to simulate a statistical data base query system and to better understand gridlock and other operational scenarios.

\subsection{3.2} {Bounding Entries in Multi-Dimensional Statistical Tables}

Approaches to statistical disclosure limitation in tabular data and to evaluating its effectiveness (\it{disclosure audit}) rely predominantly on mathematical formulations and optimization (Cox 1980, 1987, 1994, 1995; Fischetti and Salazar 2000). A promising connection between disclosure audit and Markov Chain Monte Carlo (MCMC) computation is apparent, as follows.

Consider the following problem. An NSO cannot release the internal entries of a particular multi-dimensional (n-dimensional) contingency table due to confidentiality concerns, e.g., the presence of too many small counts. It is decided to release instead the (n-1)-dimensional marginal totals, viz., obtained by adding internal entries along precisely one of the n dimensions. The NSO next must assess the adequacy of this procedure, viz., are linear estimates of disclosure cells (the small counts) sufficiently broad to permit this release? This is the \it{n-dimensional bounding problem}: determining for each internal cell exact integer linear lower and upper bounds on the cell value, given the marginals. For details of the next paragraph, see Cox (2000a).

The n-dimensional bounding problem is an integer linear programming problem and is computationally infeasible to solve except for the case of a small number of small tables. Heuristic algorithms have been offered (Buzzigoli and Giusti 1999; Fienberg 1999), all of which have been shown to be weak or to fail (Cox 2000b), for two reasons. First, sufficient conditions on the set of n (n-1)-dimensional marginal totals ensuring the existence of a feasible n-dimensional contingency table have yet to be discovered, despite five decades of research in the operations research community on what they refer to as the \it{solid transportation problem}. Heuristic algorithms based on integer operations among purported marginals will produce seemingly valid integer bounds when in fact no tables exists. Second, in three or higher dimensions, it is possible that, given integer marginals, the continuous lower or upper bound on an internal entry can be noninteger. Proposed bounding algorithms improve inexact integer lower and upper bounds using integer operations. Such methods, which proceed it\{inwards} from outside of the feasible region, are therefore incapable of crossing beyond a noninteger exact continuous bound in pursuit of the exact integer bound. Similarly, methods that rely on all-subsets analysis, e.g., as proposed at ISBA 2000 by Fienberg, are as complex computationally as a tree-structured search for feasible integer values, viz., using \it{branch and bound} methods. The n-dimensional bounding problem is daunting both theoretically and computationally.

MCMC computation within an n-dimensional contingency table is also a complex problem. Here, however, the method operates within the feasible region, using MCMC computational steps (it{moves}) defined by mathematical objects associated with algebraic geometry until recently unfamiliar in statistics (Diaconis and Sturmfels 1998). This presents two possible avenues of research. The first, and more speculative, would be to pursue exact bounds from within the feasible region, using the moves. The moves themselves are not designed to proceed to the boundary of the feasible region. However, infeasible moves are often encountered in MCMC. Instead of ignoring these moves, it may be possible to move alternately inside and outside the feasible region. By selecting moves not probabilistically but on the basis of improvement of an objective function, viz., maximizing or minimizing an internal entry, it may be possible to pursue exact bound from within the feasible region. Thus, MCMC would benefit the disclosure audit problem. Conversely, algorithms from operations research for moving around a feasible region defined by (very) many structural equations may facilitate construction of the moves and improve MCMC computation.

The second, less speculative, avenue of research would be to use MCMC to simulate the distribution of n-dimensional contingency tables consistent with the given marginals. The NSO can do so readily as it has a feasible integer starting table, viz., the original data. Using moves generated, e.g, using Gröbner bases, integer solutions are assured. It then is possible to construct, say, 95% credible regions for feasible integer values of each internal entry. These can be considered to be \it{credible exact bounds}, which in some applications may be sufficient as surrogates for exact bounds. If more than credible exact bounds are required, the credible bounds could be used to develop cut constraints for a full integer programming analysis. Note that this approach will work for the attacker only if the attacker is in possession of a feasible integer starting solution. This approach is limited by the computational feasibility of MCMC computation and Gröbner bases, both currently being actively explored.

\section {4.}{Concluding Comments}

We have described extended and new research on important problems in statistical disclosure limitation that involve Bayesian or hierarchical methods. The use of Bayesian methods to simulate disclosure attack and to assess disclosure risk appears natural. Hierarchical structure enables evaluation of the effects of measurement and sampling errors on disclosure limitation. A hierarchical Bayesian framework appears to me uniquely suited for evaluating the effects of disclosure limitation or model-generated data on data use. New research, combining Bayesian statistics and mathematical optimization, is suggested by the n-dimensional bounding problem.

\bf{Disclaimer}

The information in this article has been funded wholly or in part by the United States Environmental Protection Agency. It has been subjected to Agency review and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

\bre

\r Buzzigoli, L. and Giusti, A. (1999). An algorithm to calculate the lower and upper bounds of the elements of an array given its marginals. \it{Statistical Data Protection, Proceedings of the Conference, Lisbon, 25 to 27 March 1998}, Luxembourg: EUROSTAT, 131-147.

\r Cox, L.H. (1980). Suppression methodology and statistical disclosure control.

\jasa{75}, 377-385.

\r Cox, L.H. (1987). A constructive procedure for unbiased controlled rounding.

\jasa{82}, 520-524.

\r Cox, L.H. (1994). Matrix masking methods for disclosure limitation in microdata.

\surmeth{20}, 165-169.

\r Cox, L.H. (1995). Network models for complementary cell suppression.

\jasa{90}, 1453-1462.

\r Cox, L.H. (2000a). On properties of multi-dimensional statistical tables. Submitted: \jspi.

\r Cox, L.H. (2000b). Bounding entries in 3-dimensional transportation arrays. Manuscript.

\r Duncan, G. and Lambert, D. (1986). Disclosure-limited data dissemination (with comment).

\jasa{81}, 10-28.

\r Duncan, G. and Lambert, D. (1989). The risk of disclosure for microdata.

\jbes{7}, 207-217.

\r Federal Committee on Statistical Methodology (1994). \it {Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology}. Washington, DC: U.S. Office of Management and Budget. Available: http://www.fcsm.gov/

\r Fienberg, S. (1999). Fréchet and Bonferroni bounds for multi-way tables of counts with applications to disclosure limitation. \it{Statistical Data Protection, Proceedings of the Conference, Lisbon, 25 to 27 March 1998}, Luxembourg: EUROSTAT, 115-129.

\r Fienberg, S. and Makov, U. (1998). Confidentiality, uniqueness, and disclosure limitation for categorical data.

\jos{14}, 385-397.

\r Fischetti, M. and Salazar-Gonzales, J. (2000). Models and algorithms for optimizing cell suppression in tabular data with linear constraints.

\jasa{95}, to appear.

\r Fuller, W. (1993). Masking procedures for microdata disclosure limitation.

\jos{9}, 383-406.

\r Lambert, D. (1993). Measures of disclosure risk and harm.

\jos{9}, 313-331.

\r Little, R. (1993). Statistical analysis of masked data.

\jos{9}, 407-426.

\r Paass, G. (1985). Disclosure risk and disclosure avoidance for microdata.

\jbes{6}, 487-500.

\r Rubin, D. (1993). Discussion: statistical disclosure limitation.

\jos{9}, 461-468.

\r Samuels, S. (1998). A Bayesian, species-sampling approach to the uniques problem in microdata disclosure risk assessment.

\jos{14}, 373-383.

\r Skinner, C. and Holmes, D. (1998). Estimating the re-identification risk per record to microdata.

\jos{14}, 361-372.

\ere

\end{document}

NERL-RTP-IO-00-169

TECHNICAL REPORT DATA

1. REPORT NO. EPA/600/A-00/082	2.	3.1
4. TITLE AND SUBTITLE Towards a Bayesian Perspective on Statistical Disclosure Limitation		5. REPORT DATE
		6.PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Lawrence H. Cox		8.PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS National Exposure Research Laboratory, Research Triangle Park, NC		10.PROGRAM ELEMENT NO.
		11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS NATIONAL EXPOSURE RESEARCH LABORATORY OFFICE OF RESEARCH AND DEVELOPMENT U.S. ENVIRONMENTAL PROTECTION AGENCY RESEARCH TRIANGLE PARK, NC 27711		13.TYPE OF REPORT AND PERIOD COVERED
		14. SPONSORING AGENCY CODE USEPA

15. SUPPLEMENTARY NOTES

16. ABSTRACT

National statistical offices and other organizations collect data on individual subjects (persons, businesses, organizations), typically while assuring the subject that data pertaining to them will be held confidential. These data provide the raw material for statistical data products (tabular summaries, microdata files comprising data records pertaining to individual subjects, and, potentially, public statistical data bases and statistical query systems) which the statistical office disseminates to multiple, broad user communities. Statistical disclosure limitation (SDL) refers to the problem and methods for thwarting re-identification of a subject and divulging the subject's confidential data through analysis or manipulation of disseminated data products. SDL methods abbreviate or modify the data product sufficiently to thwart disclosure. SDL problems are typically computationally demanding; several have been shown to be NP-hard. Many SDL methods draw upon statistical, mathematical or optimization theory, but at the same time heuristic and partial approaches abound. Contributions from a Bayesian perspective have been few but are increasing. A strong theoretical connection between definitions of statistical disclosure, measurement of disclosure risk, and evaluation of SDL methods is lacking. This suggests opportunities for Bayesian and hierarchical approaches. Selected opportunities and associated SDL methodological issues are discussed.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b.IDENTIFIERS/ OPEN ENDED TERMS	c.COSATI

18. DISTRIBUTION STATEMENT RELEASE TO PUBLIC	19. SECURITY CLASS *(This Report)* UNCLASSIFIED	21.NO. OF PAGES 6
	20. SECURITY CLASS *(This Page)* UNCLASSIFIED	22. PRICE