DISCUSSION

Nancy L. Spruill, Office of the Assistant Secretary of Defense
for Force Management & Personnel
Pentagon, Rm. 4C761, Washington, D. C. 20301

Introduction

I really enjoyed reviewing the four papers given in this session. These papers include some of the best written, clearly laid out ones that I've read. They do a good job of discussing the complex issues in privacy protection and disclosure avoidance. They also present some good new ideas for avoiding disclosure and enhancing data analysis when compared to existing techniques.

I just want to quickly review each paper and then ask a question of each of the presenters to begin the discussion.

Paper#1

Disclosure Avoidance Techniques in the Canadian Censuses of Population and Agriculture by Mary J. March and Douglas A. Norris

This paper is clearly written and looks at disclosure issues for the Census of Population -- where small frequency is the main problem -- and the Census of Agriculture -- where the problem comes from one or two respondents contributing almost all the information in a cell.

The Census of Population provides data both in tables and in building blocks so that the user can produce tables. The masking techniques used are random rounding to the base 5 and suppression. The Census of Agriculture provides data in tables but does not accomodate specific user requests. The masking technique is a customized system of cell suppression (and complementary cell suppression) because of small sample sizes or one or two dominate farms.

QUESTION: Suppose a data intruder gets out of George and Diane's paper and slips into Canada. Can this intruder design multi-custom tables for data from the Census of Population that will lead to disclosure problems not seen in the individual tables? How do your disclosure avoidance techniques of random rounding and suppression protect against multi-custom tables "designed" by an intruder?

Paper#2

The Risk of Disclosure for Microdata by George Duncan and Diane Lambert

This is a good, well-written paper that analyzes the risk of disclosure for several cases. It looks at two kinds of disclosure -- identity and attribute -- and two kinds of data "intruders" -- an uninformed outsider and an informed insider. A unique and useful thing about this paper is the introduction of a loss function that describes the intruders goals with respect to the data. With a known loss function, the data releaser can test the amount of risk for a proposed data release. Also, he or she can modify the masking techniques to minimize the risk.

I especially appreciated the numerical examples given in this paper. See page 12 for a simple example. I had a much better understanding of the issues involved from these examples and from the discussion of what a data releaser can do to "dissuade linking" and hence to dissuade disclosure. I found this discussion very informative.

What I'd like to see is more examples looking at different loss functions. These examples should include

1) translating the results for the data releaser to what he or she can do to provide protection and

2) translating the results for the data user on how he or she can get better information when doing analyses using the released data.

QUESTION: Each data intruder has his or her own loss function. How do the data releasors provide protection again all of these threats?

Paper#3

Further Development of the Randomized Response Technique for Masking Dichotomous Variables by Jay Kim

This paper is well written and examines the masking technique of randomized response for data taking the values 0 or 1. It shows that this technique can preserve -- either exactly or by dividing by a constant -- the correlation structure of the unmasked data with any other data (those taking 0 or 1 values or those taking continuous values). The techniques in this paper are still evolving and developing. But, I think they have real promise. One of the biggest problems in masking data is masking zero values. Many masking techniques that work well for continuous data, destroy important information for discrete data. This technique includes zero values for the user while still preserving the correlation structure and hence is of great value to data users.

Of course, we're all waiting to see the technique expanded to masking of multi-level discrete data, to masking continuous data, and to masking mixed data.

QUESTION: What are the problems with expanding your technique to multi-level discrete data? Continuous data and discrete data really do pose different problems for data masker and data user. Does your technique have promise for continuous data?

Paper#4

Assessing Quality of Randomized Response: Were Instructions Followed? by James Schmeidler

This paper addresses a slightly different problem. In the previous three papers the data collector knew the true value of the data and was trying to protect the identity of the person who gave him or her that data. In this paper, the data collector doesn't have the real stuff. Only the person providing the data does and he or she may not want even the data collector to know the true value. Statistically we can get around this problem, but only if the person with the data trusts us and "plays by the rules." This paper looks at how we could test the hypothesis of "playing by the rules."

I had a little trouble keeping straight the terms validity, reliability, unreliability, etc. If I could see it in terms of Xs, Ys, αs , and βs , even if I had to go to an appendix, I would have a better understanding. However, after I got into the discussion and application of randomized response, I did better.

The paper discusses a pilot telephone study that used 3 flips of a coin to tell the respondent how to answer the questions (yes if all heads, no if all tails, and the truth otherwise). The study asked 2 questions about drugs -- one about lifetime use and the other about recency of use. The results (interpreted loosely) were that the randomized response directions were followed on the two questions.

QUESTION: If the test showed that the directions were likely not followed, could the survey takers salvage any results? Also, suppose you got results from a truely anonimous survey (one that comes in the mail and has no markings on the return envelope or survey form) that didn't involve randomized response. And you also got results from another telephone survey that used randomized response. Could you compare these results and have another way to get at whether those participating in your randomized response survey were "playing by the rules?"