

Discussion of Papers on Disclosure Control and Data Access

Stephen E. Fienberg

**Department of Statistics & Center for
Automated Learning and Discovery
Carnegie Mellon University**

Background

- **Tension between confidentiality restrictions and data access:**
 - new statistical techniques ==> release of altered data**
- **Expanded modes of dissemination:**
 - PUMS.
 - WWW-accessible files.
 - Links to other data bases.
- **Concerns re growth of datamining and private uses of public data.**

Buzzigoli-Biggieri

- **Are safe settings viable alternatives?**
 - Remote execution of statistical programs.
 - Luxembourg Income study
 - Licensing, e.g.,
 - Health Retirement Survey/ U. Michigan.
 - Intermediaries, e.g.,
 - Statistics Canada and Data Liberation Initiative.
 - Special, restricted access data centers, e.g.,
 - U.S. Census Bureau Data Centers.

Buzzigoli-Biggieri

- **Link to GIS systems?**
- **Expanded geographic information is biggest threat to confidentiality:**
 - Issue re U.S. census PUMS and other releases.
 - **Anderson & Fienberg, 2001 (last session)**

Takemura

- **Per-record identification risk assessment:**
 - In tradition of Skinner-Holmes and Fienberg-Makov.
 - Use of Lancaster additive model for contingency tables instead of log-linear or logisitic models.
- **Example: $14 \times 2 \times 91 \times 5 \times 14 \times 7 \times 2 \times 5 \times 2 \times 10$ sparse table from 1990 Census PUMS:**
 - Lancaster model has implementation problems.
 - Estimated “rate of population uniques” is low, 1/10.
 - Example actually ideal for log-linear model analysis.

Karr-Sanil

- **Web-based geographically-aggregated data release:**
 - Uses greedy-search algorithm to collapse units.
 - System “actually works” and has led to release of previously restricted data.
- **General table server:**
 - Innovative system design.
 - New statistical criteria and methods for assessing risk and releasing data --currently working with 2^{16} table from National Long Term Care Survey.

Safe Data vs. Safe Settings

- **The only *useful* public data are *released* public data!**
 - Safe setting approach can only work as a stop-gap measure.
- **We are making important progress on technical solutions to disclosure limitation that should increase data access in future.**