

THE NISS

DIGITAL GOVERNMENT PROJECT

Progress Report: April 24, 2000

Alan F. Karr

karr@niss.org

Outline

Introduction: Alan Karr

Geographical Aggregation: Ashish Sanil

Table Server: Alan Karr

NASS Prototype: Ashish Sanil

Other Activities: Alan Karr

The Research Team

NISS: Alan Karr, Ashish Sanil, Jaeyong Lee [, James Hilden–Minton]

CMU: Adrian Dobro, George Duncan, Stephen Fienberg, Latanya Sweeney

LANL: Sallie Keller–McNulty

MCNC: Bonnie Parrish, Karen Litwin, Syam Sundar, ...

Review

Build a Web-based query system that

1. Is dynamic and *history-dependent*
2. Dispenses statistical analyses rather than (micro)data
3. Uses statistical technology to preserve confidentiality
4. Reflects user community needs

Implement the system on “live” Federal agency databases

Understand how the system is used and performs

Evaluate disclosure risk models and risk reduction strategies at *realistic* scales, using the system as testbed

Summary of Progress to Date

- Algorithms for geographic (or other) aggregation (Sanil, Lee, Karr)
- Statistical implications of aggregation (Lee, Sanil, Karr)
- Prototype table server design (Karr, Sanil, Hilden–Minton)
- QHDB schema for table server (Sanil, Karr, Hilden–Minton)
- NASS prototype under construction (Karr, Lee, Sanil, MCNC)
- Scalability of methods to compute bounds (Fienberg, Dobro)
- Bayesian framework for confidentiality protection (Duncan, Keller–McNulty)
- Confidentiality Reading Group, involving NISS, RTI, other Research Triangle individuals
- Interactions with other DG projects (Columbia, UNC) and other NISS IT initiatives

Table Server Prototype

Data: Sample Census data set with

- 48,824 cases
- 8 (after trimming) categorical variables: Age, Education, Employer type, Marital status, Race, Salary, Sex, Work Hours

Query: Sub-table of full 8-way table

Response: Requested sub-table (FTP, character display, visualization) or statement that it cannot be released

Problem Conceptualization

Based on partial ordering of tables

- Core releases
- Direct releases
- Indirect releases
- Released/unreleased frontier

New Release \equiv Movement of Frontier

- How far?
- How often?
- By whom?

Risk Criteria

- Predictive capability for sensitive variable
- Accuracy of IPF reconstruction of full table
- [Accuracy of LP bounds on cell entries]
- [Entropy]

System Design

- Flow chart
- QHDB schema
- Visualization as a means of risk reduction
- Visual interfaces incorporating association

Fienberg/Dobro

Progress: Formal results on bounds for tables and their relationship to log-linear model and graphical structures. New theorems for the "decomposable case" and extensions that reduce the bounding problem to smaller dimensional components.

With Duncan, exploration of formal structures required to weight the tradeoff between disclosure risk and societal gains from data release, using a formal Bayesian information theoretic approach.

Current Challenges: Scaling up the results so that they are computationally feasible for actual government survey settings.

Products: Papers (PNAS); code to be incorporated in table server.

Duncan/Keller–McNulty

Progress: Initial steps toward formal Bayesian decision–theoretic framework for confidentiality protection through disclosure limitation. The framework explicitly incorporates disclosure risk and data utility. It also permits the comparison of disclosure limitation through matrix masking and generation of synthetic data.

With Fienberg, exploration of formal structures required to weight the tradeoff between disclosure risk and societal gains from data release, using a formal Bayesian information theoretic approach.

Current Challenges: Formally analyze the impact on disclosure risk and data utility of data swapping. Better understand synthetic data as a disclosure limitation tool. Develop associated procedures for disclosure risk estimation and disclosure limitation that scale.

Products: New algorithms. Review paper on confidentiality and disclosure limitation, to be published in the International Encyclopedia of the Social and Behavioral Sciences (Duncan).

The Next Six Months

- Project Web site
- Complete NASS prototype; write associated paper(s).
- Functional table server prototype with dynamic risk estimation and visualizations. *Major scalability questions will remain.*
- Initial concepts of query, risk, response for regression server.
- [Initial consideration of longitudinal data]