THE NISS
DIGITAL GOVERNMENT PROJECT

Overview and Highlights

Alan F. Karr

karr@niss.org

dg.o

May 16, 2000

Project Goals

Build Web-based query systems that

1. Dispense statistical analyses rather than (transformed) microdata
2. Are dynamic, with *history-dependent* assessment of disclosure risk
3. Use statistical technology to ensure confidentiality
4. Reflect user community needs

Understand how the systems are used and perform

Evaluate disclosure risk models and risk reduction strategies at *realistic* scales, using the systems as testbeds

Implement (ultimately) the systems on "live" Federal agency databases

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, . . .

Summary of Progress

- Algorithms for geographic (and 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, [Fienberg,] Keller–McNulty)
- Confidentiality Reading Group, involving NISS, RTI, other Research Triangle individuals

Geographic Aggregation: NASS Setting

Data: Survey of farms for fertilizer/pesticide usage, by crop, chemical and year.

Data Table: Has columns

[FarmID, Crop, Chem, Year, Location_Code, Acres, Amount_Applied]

Query: Request for lbs./acre of (Chemical = Chem_X) applied to (Crop = Crop_Y) in (Year = Year_Z) averaged over all farms at (Location = Location_L) (that applied Chem_X to Crop_Y in Year_Z).

Response: Application/acre averaged over all farms satisfying the query conditions

Disclosure Criteria

For a value to be releasable,

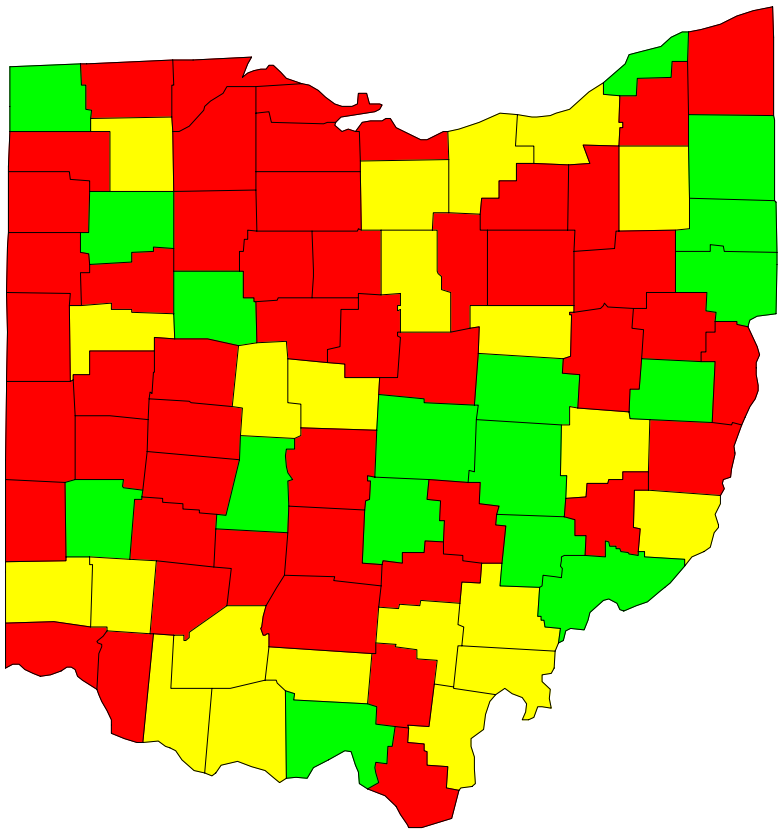
- The number of farms must be ≥ 3
- No farm satisfying the query conditions can contain more than 60% of the total acreage

This is a severe restriction. For corn in Ohio, of 88 counties, 39 are disclosable and 49 undisclosable

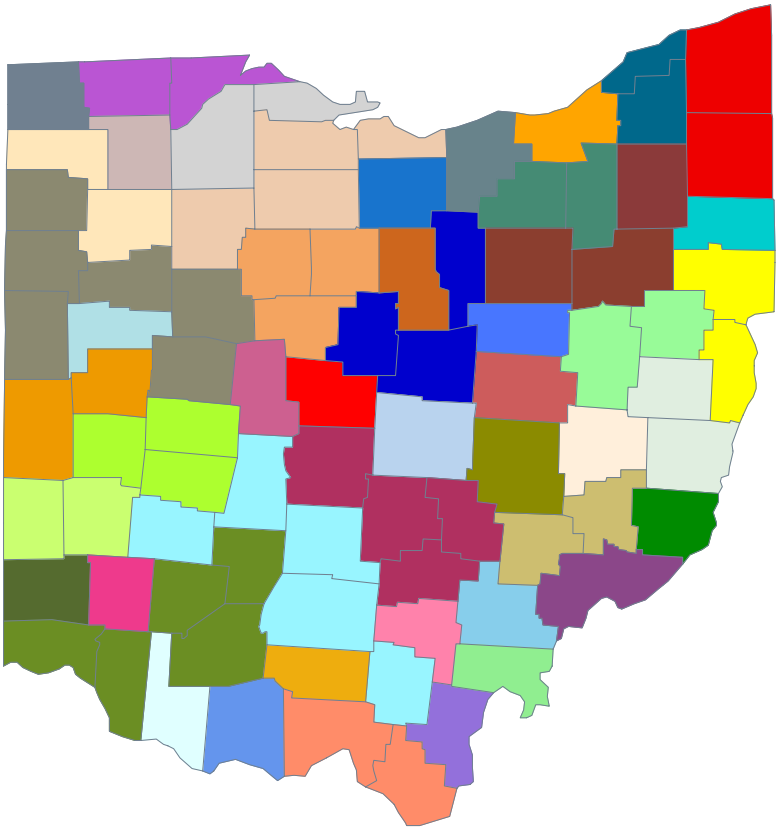
Approach: Geographic Aggregation

- Aggregate counties into disclosable "super–counties"
- Criteria: Purity, smallness, compactness
- Methods (must be automatic and fast)
  - Heuristic: Visit each undisclosable (super) county in random order and merge with a neighboring (super–) county until only disclosable (super–) counties remain. Purity, smallness and hybrid versions.
  - Simulated annealing (with aggregations represented as bit strings) using various objective functions: smallness, compactness

Disclosability

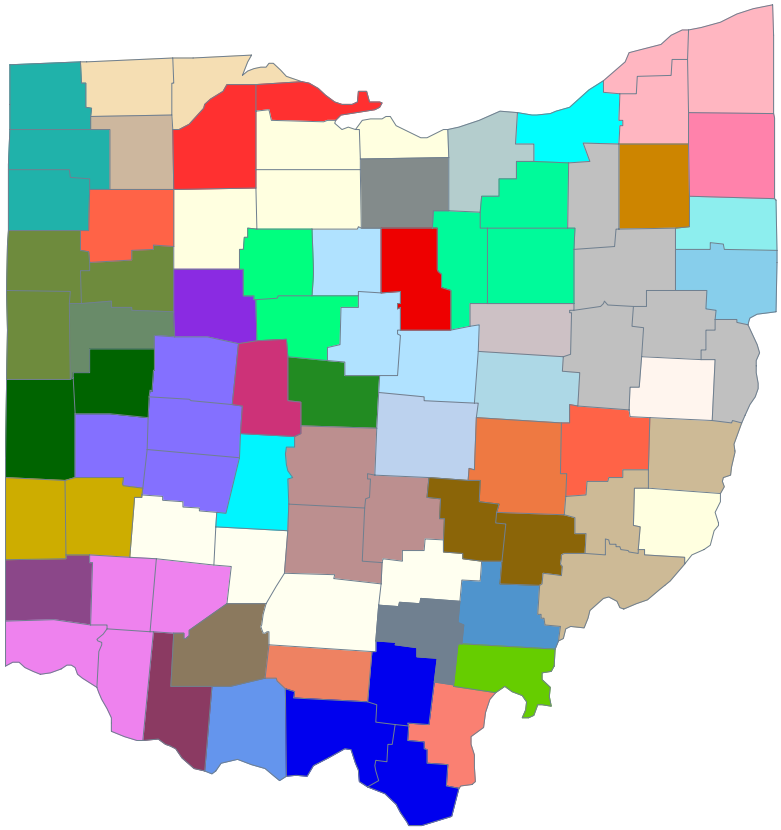


Heuristic – Smallness

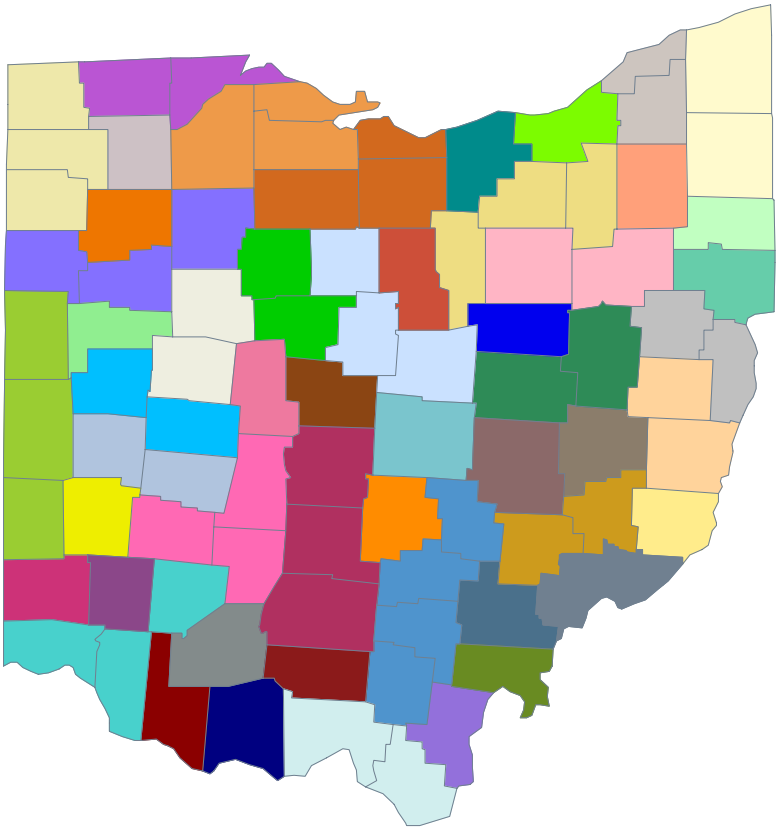


Heuristic – Purity

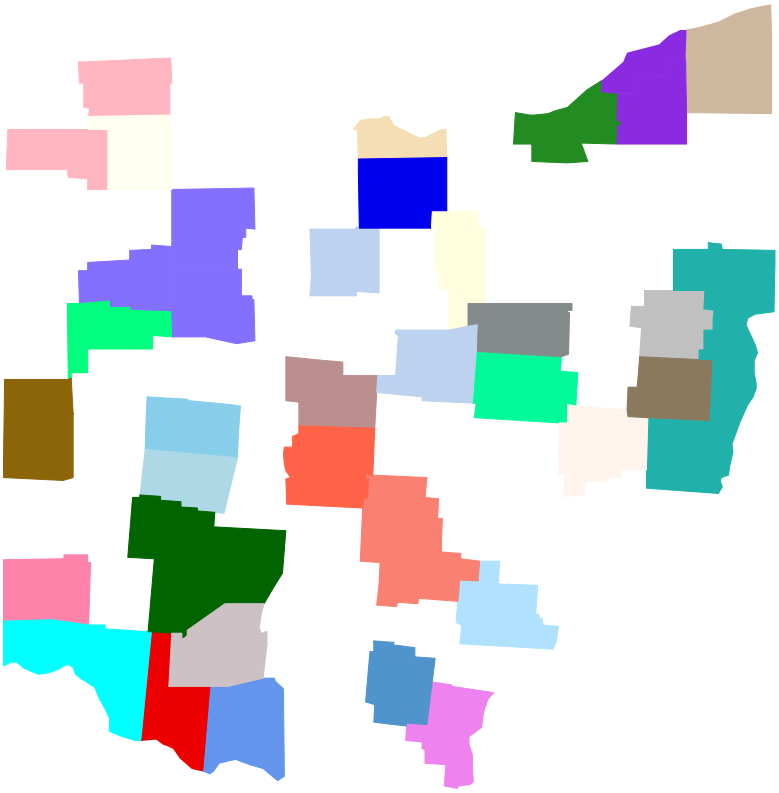
Simulated Annealing – Smallness



Simulated Annealing – Strong Smallness



Simulated Annealing – Compactness



Summary of Results

Method	#(SC)
Heuristic – pure	52
Heuristic – small	49
SA – small	50
SA – strong small	51
SA – compactness	50

Table Server Prototype

Data: Sample Census data set with

- 8 (after trimming) categorical variables: Age, Education, Employer type, Marital status, Race, Salary, Sex, Work Hours
- 48,842 cases; 2880 cells, of which 1695 are non-zero; maximum cell count = 1255

Query: Sub-table (cross-tabulation) of full 8-way table

Response: One of

- Requested sub-table (download, character display, visualization)
- Statement that it cannot be released
- Requested sub-table to which risk reduction strategies (e.g., swapping, aggregation) have been applied

Problem Conceptualization

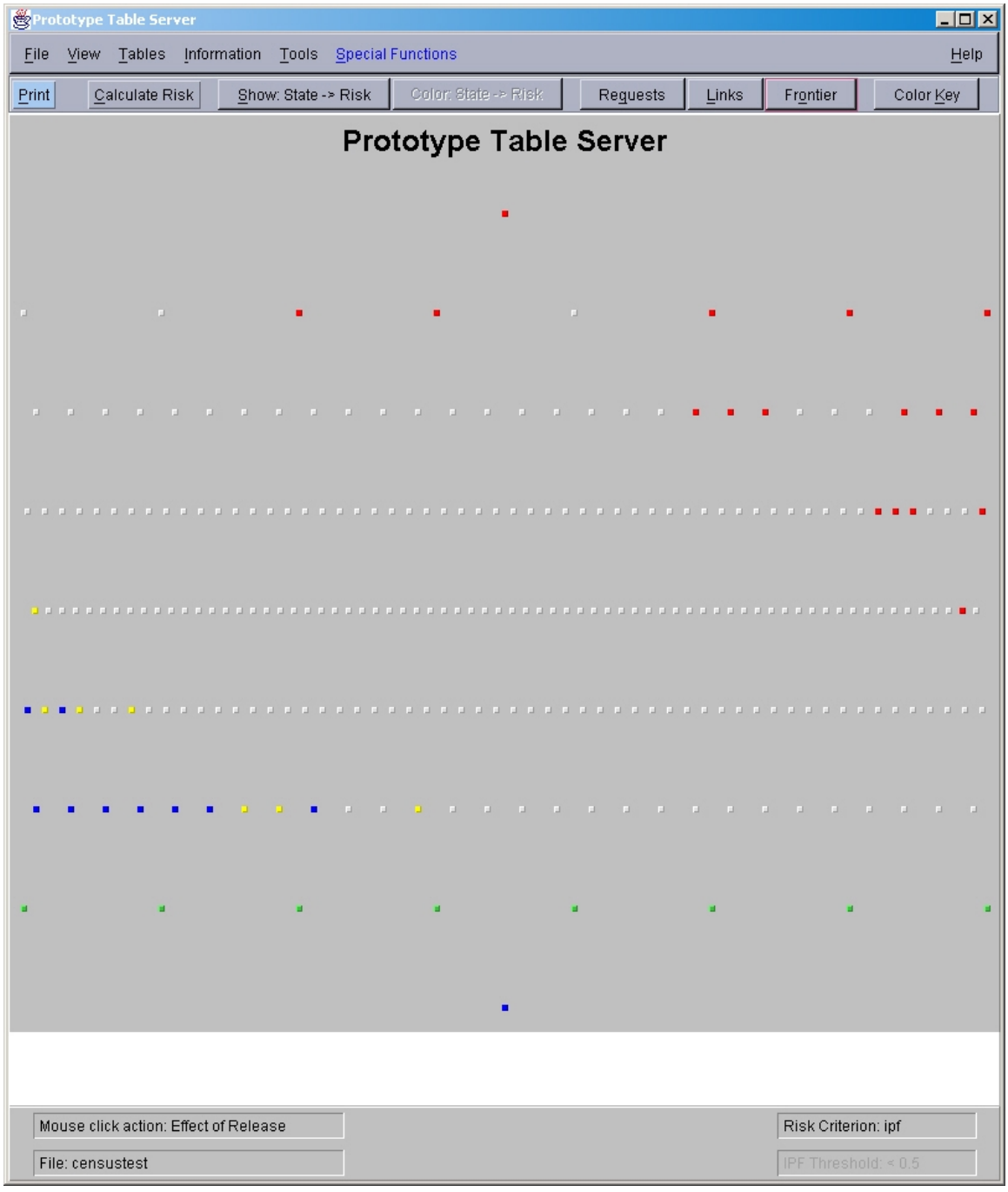
Based on partial ordering of tables

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

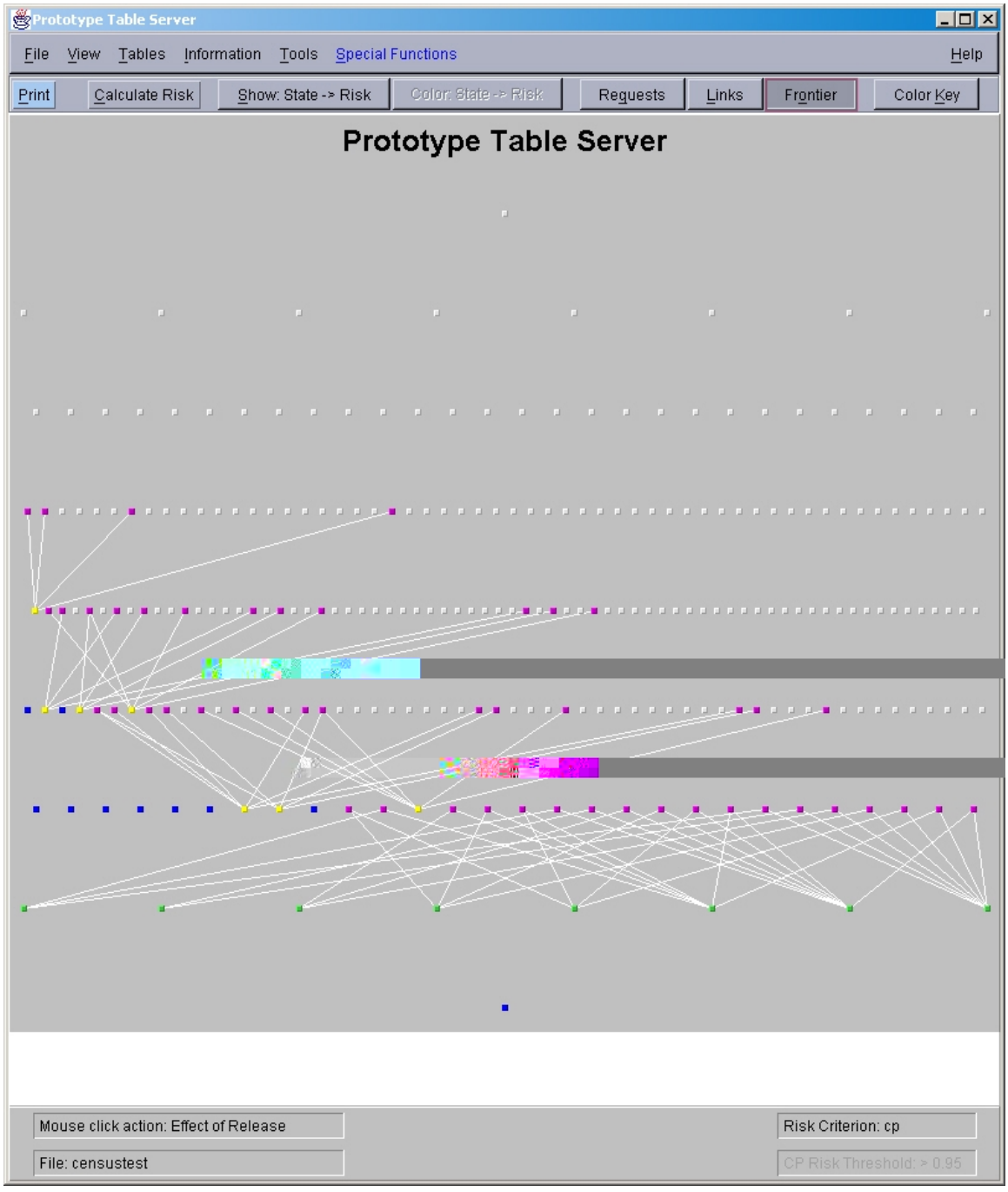
New Release ≡ Movement of Frontier

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

The Query Space



The Unreleased Frontier



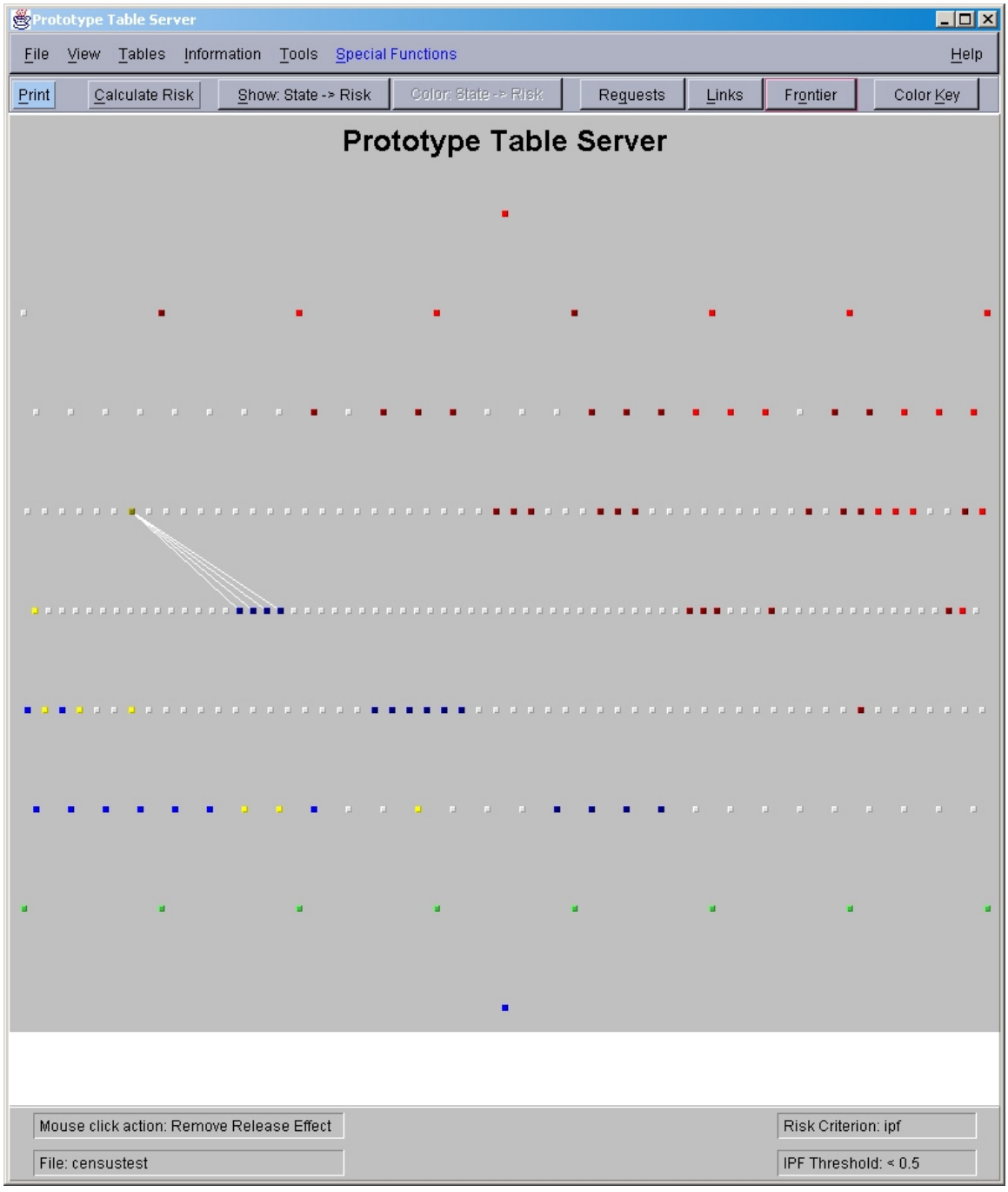
Risk Criteria

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

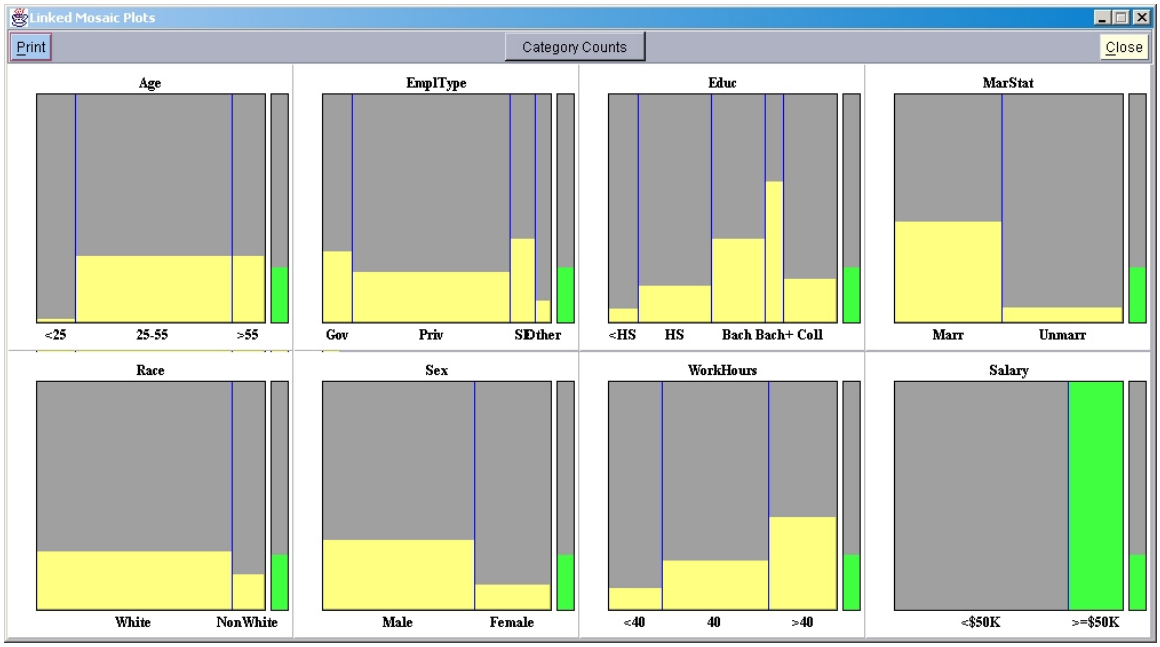
System Operator–Centric Capabilities

- Representation and visualization of query space
- Evaluation of the effect of requested releases
- Visualization as a means of risk reduction. Example: Association among variables

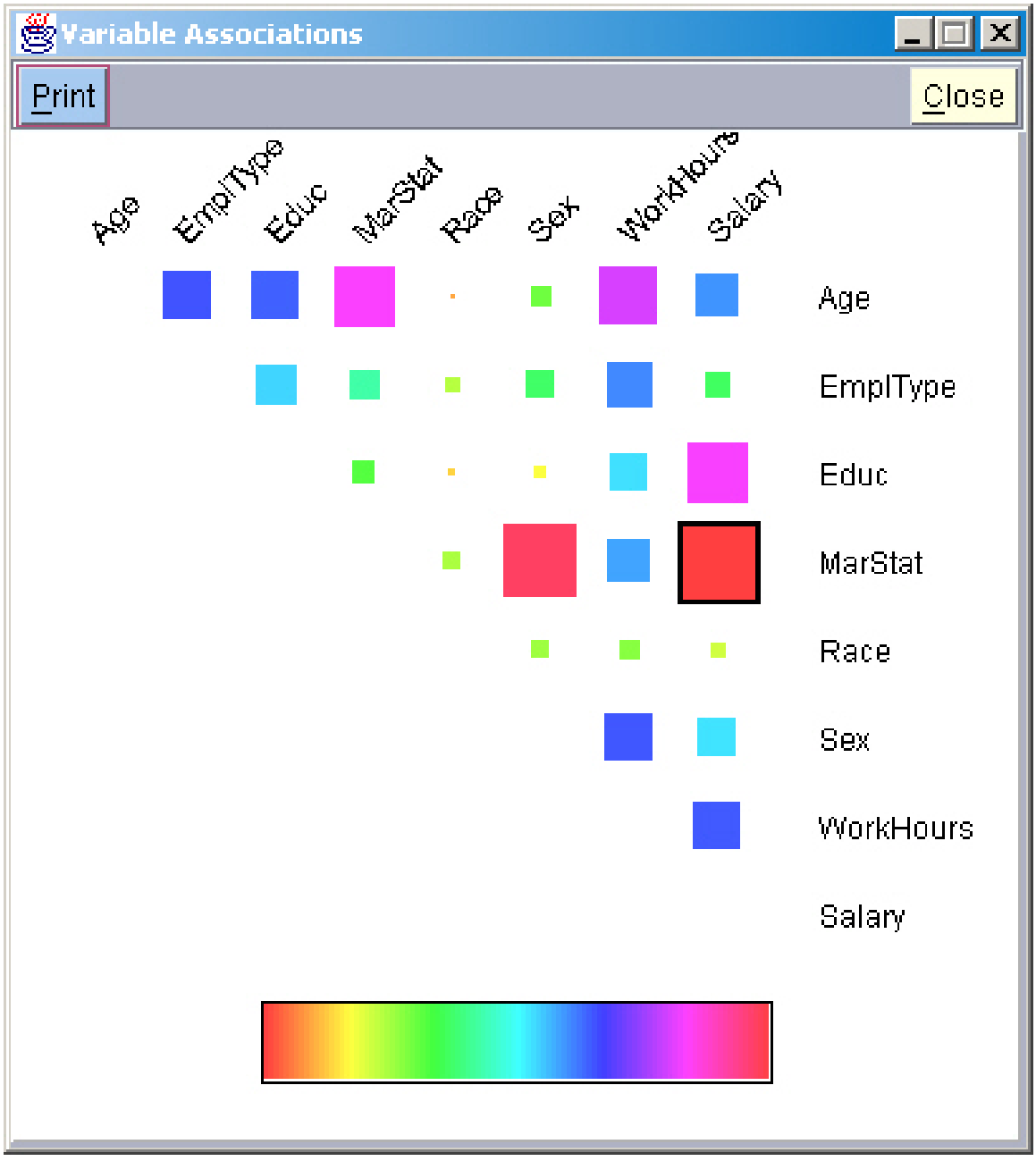
Effect of a Release



Linked Mosaic Plots



Association Coefficients



The Next Six Months

- Project Web site: `www.niss.org/dg`
- Complete NASS prototype
- Functional table server prototype with dynamic risk estimation
- Initial concepts of query, risk, response for regression server
- [Initial consideration of longitudinal data]

Longer–Term Issues

- Scalability of
  - Methods to evaluate and reduce disclosure risk
  - Techniques to represent and visualize query space
- User issues, especially movement of frontier
- Inclusion of the *value* of releasing information