THE NISS
DIGITAL GOVERNMENT PROJECT

Alan F. Karr   Sallie Keller–McNulty
karr@niss.org   sallie@lanl.gov

October 19, 2000

Outline

- Introduction to NISS
- Introduction to the Project
- Geographical Aggregation
- Table Servers

NISS

- Research institute, established in 1990 to enlarge the future of statistics, by identifying, catalyzing and fostering high-impact cross-disciplinary research involving the statistical sciences
- Located in Research Triangle Park, North Carolina; sponsored by 5 statistical societies and 5 NC organizations
- Work carried out in collaborative, cross-disciplinary (often, geographically distributed) projects

  Project areas: environment, computer network intrusion, ISP customer churn, drug design, education, gene expression, software engineering, transportation, materials science, information technology (Digital Government)

  Personnel: NISS leadership; senior personnel (from universities, corporations, national laboratories, government agencies; postdoctoral fellows (2–3 year appointments); graduate students (with their advisors or as interns)
- Affiliates program, with 16 corporations, 9 government agencies and national laboratories and 16 university departments, focuses on emerging areas and informs project development

See `www.niss.org`

Project Goals

Build Web-based query systems that

1. Disseminate statistical analyses rather than (transformed, altered, synthesized, . . .) microdata
2. Are dynamic, with *history-dependent* assessment of disclosure risk for each query
3. As a result, reflect user community needs: data are probed more deeply in regions of user interest

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

Implement (ultimately) the systems on Federal agency databases, and

Understand how the systems are used and perform

See `www.niss.org/dg`

The Current Research Team

NISS: Alan Karr, Ashish Sanil [, Jaeyong Lee, Karen Brady, Christopher Holloman]

Carnegie Mellon University: Adrian Dobra, George Duncan, Stephen Fienberg, Andrew Moore, Stephen Roehrig, Mario Trottini

Los Alamos National Laboratory: Sallie Keller–McNulty

MCNC: Joel Hernandez, Sousan Karimi, Karen Litwin, Syam Sundar

Ohio State University: Alan Saalfeld

Partner Agencies: Bureau of Labor Statistics, Census Bureau, National Agricultural Statistics Service, National Center for Education Statistics, National Center for Health Statistics

Thrusts of the Research to Date

- ❇ Algorithms for geographic aggregation, incorporated in NASS prototype
- Statistical implications of aggregation
- ❇ Table servers: prototype and preliminary design specifications
- Scalable methods to compute bounds for table entries
- Bayesian framework for confidentiality protection, accounting for the *value* as well as the risk of releasing information

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, County, Acres, ApplicationRate]

Query: For application rate of Chem = X applied to Crop = Y in Year = Z for farms in Location = L (that applied Chem = X to Crop = Y in Year = Z)

Response: Application rate averaged over all farms (weighted by size) satisfying the query conditions, *provided* ...

Release Rule: For the application rate in a unit to be disclosable, (1) The number of farms must be ≥ 3, *and* (2) No farm satisfying the query conditions can contain more than 60% of the total acreage.

Approach: Geographic Aggregation

Aggregate counties into disclosable “super–counties,” using various criteria:

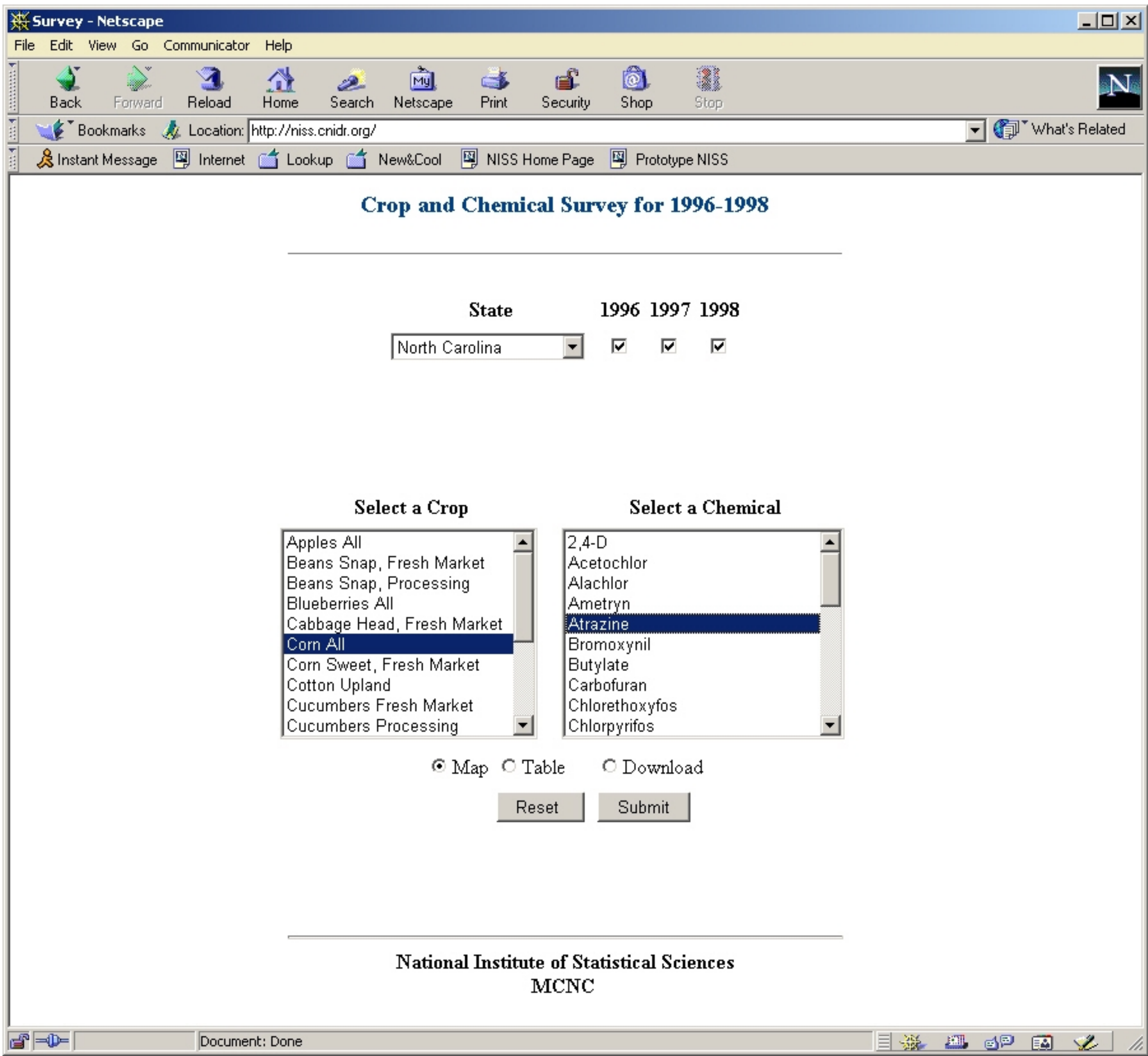
- Purity (of data from disclosable counties)
- Smallness (of supercounties)
- Compactness (of supercounties)

Heuristic methods (automatic and fast): Examine each undisclosable (super) county in random order and merge with a neighboring (super–) county until only disclosable (super–) counties remain:

- Purity
- Smallness
- Multi-step: S then P

Can also use simulated annealing (with objective functions such as compactness), but too slow

Input Screen



See `niss.cnidr.org`

Map Output

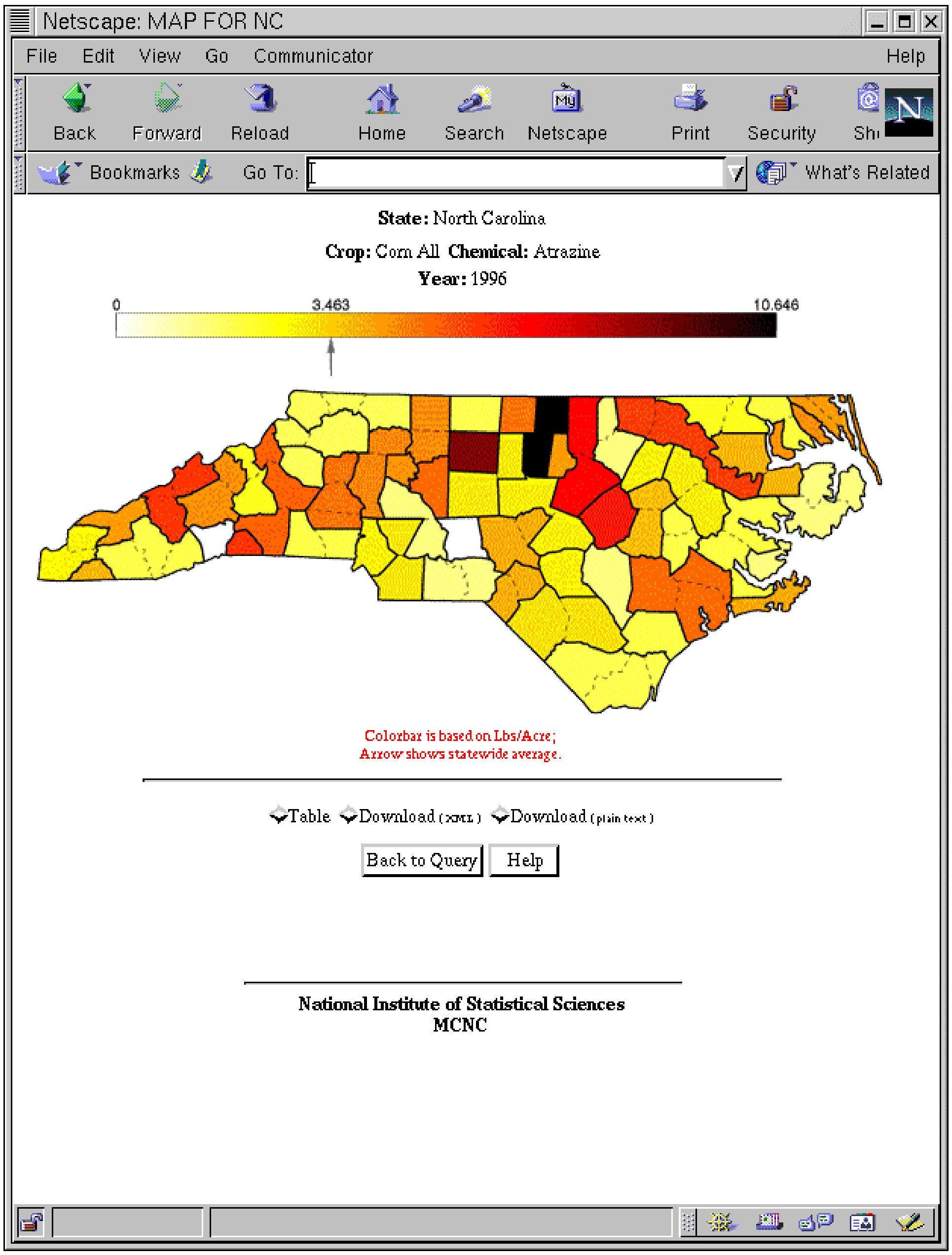

Netscape: MAP FOR NC
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shi
Bookmarks Go To: What's Related
State: North Carolina
Crop: Corn All Chemical: Atrazine
Year: 1996
0
3.463
10.646
Colorbar is based on Lbs/Acre;
Arrow shows statewide average.
Table Download (XML) Download (plain text)
Back to Query Help
National Institute of Statistical Sciences
MCNC

Table Servers

Database: Large (40 variables, 2^{40} cells) contingency table, containing either counts or totals

Query: Sub-table (cross-tabulation) of "main" table

Response: One of:

- Requested sub-table
  - XML download
  - HTML screen display
  - Visualization
- Statement that the requested sub-table cannot be released (Is this too informative?)
- Requested sub-table to which *risk reduction strategies* (e.g., cell suppression, swapping, aggregation, jittering) have been applied

Prototype Table Server

1993 Current Population Survey (CPS) data set with:

- 8 categorical variables: Age, Education level, Employer type (e.g., private sector), Marital status, Race, Salary, Sex, Work Hours (previous week)
- 48,842 cases
- 2880 cells, of which 1695 are non-zero; maximum cell count = 1255

Risk Criteria:

- Accuracy of IPF reconstruction of full table
- [Predictive capability for Salary]

Software: Java Swing application ($\sim$ 4,000 lines of code)

Problem Conceptualization – 1

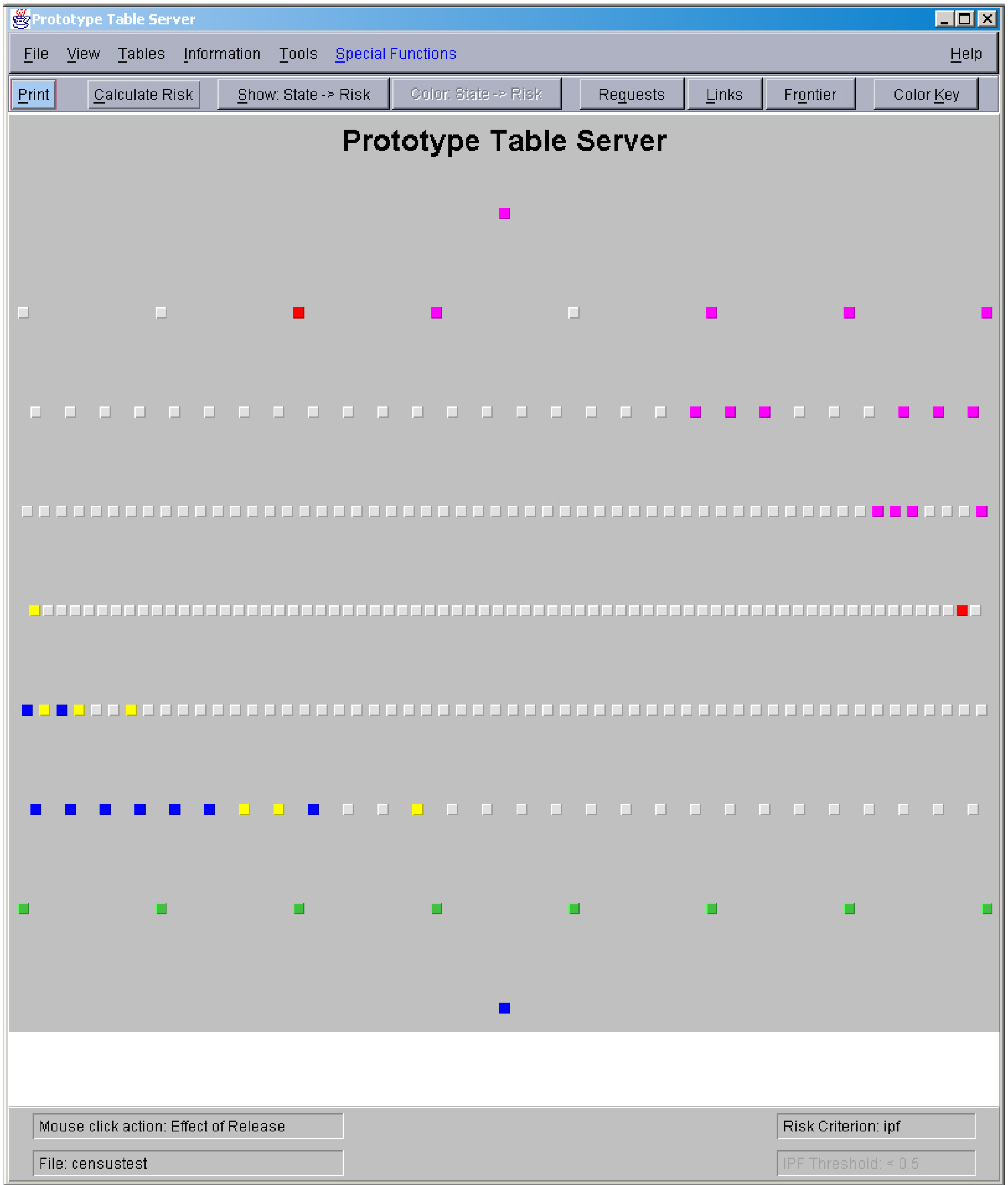
Query Space $\mathcal{Q}$ = set of all sub-tables, partially ordered by set inclusion of variables

System State specified by

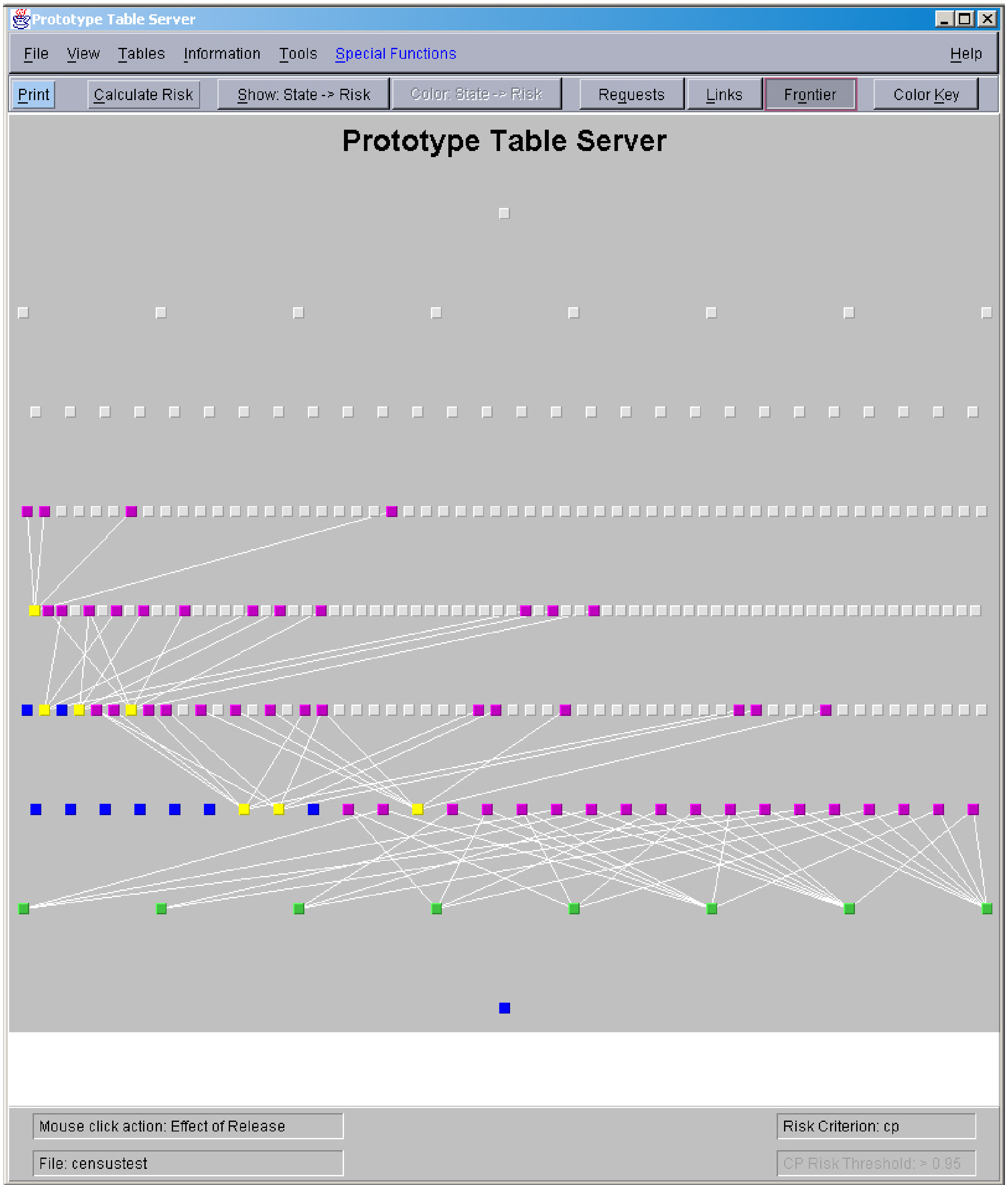
- *Core releases* when the system begins operation
- *Direct releases* in response to user queries
- *Indirect releases* — unrequested children of direct releases
- *Frontier* of released sub-tables
- *Eligibility* for release (example: one step above the frontier)
- *Unreleasable* tables whose release would cause system risk to become too high

New release $\equiv$ movement of frontier

The Query Space



Tables Eligible for Release



Problem Conceptualization – 2

Risk measured by a function $\mathrm{RiskFn}(\mathcal{R})$, where $\mathcal{R}$ is any subset of $\mathcal{Q}$ (corresponding to the current set of direct, indirect and core releases)

- RiskFn must be monotone with respect to the partial ordering (hence $\mathrm{RiskFn}(\mathcal{R})$ depends only on the released frontier $\mathrm{RelFron}(\mathcal{R})$)
- “Too risky” means

$$\mathrm{RiskFn}(\mathcal{R}) > \alpha,$$

  where α is a system–operator–set threshold

Release rules that determine which of several subtables requested for release will be released. For example, rules can account for which other tables become too risky, or the *value* of releases.

Release Rules: What Becomes Too Risky?