NISS

National Institute of Statistical Sciences
Research Triangle Park, NC 27709

Information Seminar on Statistical Disclosure Limitation

December 7, 2006

Goals

- Introduce fundamental problems and methods of statistical disclosure limitation (SDL)
- Present concrete examples
- Stimulate questions and discussion

Program

9:30 AM	Introduction, Risk-Utility Formulations, Data Swapping Alan Karr, NISS
11:00	Tabular Data George Duncan, Carnegie Mellon University
12:00 N	Lunch (on your own)
1:00 PM	Remote Access Servers: Alan Karr
1:30 PM	Synthetic Data and Related Topics Jerome Reiter, Duke University
2:30 PM	Break
3:00 PM	Research Frontiers: Alan Karr
4:00 PM	Open Discussion

Introduction

Some Terminology

- Data: flat file of subject-indexed records (rows) containing attributes (columns) of the subjects
  - Attributes may be categorical or numerical
  - Ignores a lot of other things that really are data: Images, sound, video, free-form text, …
  - Ignores relational databases
- Data owner (federal statistical agency)
- Legitimate user
- Intruder

The Fundamental Issue: Agencies Make Tradeoffs Between

- Minimizing disclosure risk
  - Mandated by law: protect data subjects' privacy
  - To maintain quality
- Maximizing data utility
  - Policy
  - Research, especially statistical inference

SDL

"Do something to keep data warehouses from becoming data cemeteries"

Forms of Disclosure

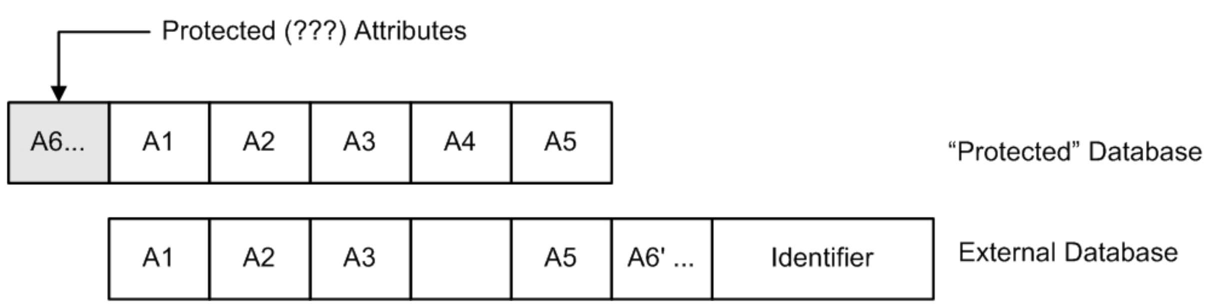
- Identity disclosure
  - Record is associated with a particular subject
- Attribute disclosure
  - Value of sensitive attribute is disclosed, with or without identity disclosure
- Inferential disclosure
  - Identity or attribute disclosure on a statistical basis
- False positive
  - Intruder acts on basis of incorrect information

Identity Disclosure

- Possible via
  - Explicit identifier: name and address, SSN, …
  - Implicit identifier: “Occupation = Mayor of New York”
  - Extreme data values: Income = $\$10^9$
  - Rare attribute combinations: State = ND, Ethnicity = Korean, Age = 50, Gender = F, NumberChildren = 8, Occupation = Statistician
  - “Recognition”
  - Linkage to external database

Record Linkage

- Locate external database containing
  - Attributes also in the released database
  - Identifiers
- Match records using common attributes



Attribute Disclosure

- Example: small count cells in categorical data
- Suppose
  - Data contain average income by (Age,Race,ZIP)
  - Only two people with
    - Age = 56-60
    - Race = white
    - ZIP = 27709
  - I am one of those two
- Then I know the income of the other one

Inferential Disclosure

- Example: income can be predicted reliably from other attributes
- Involves uncertainty
- Can be “incidental”
  - In a regression of income on age, some values lie “on” the regression line

How Easy is It?

- Most people can be identified by
  - Date of birth (MM/DD/YYYY)
  - Gender
  - 5-digit ZIP code
- Finding these items on the web is very easy
  - Voter records
  - Property tax records
  - ChoicePoint

Finding ZIP Code and Gender


NORTH CAROLINA STATE BOARD OF ELECTIONS

SBOE Home :: Campaign Finance :: En Español :: Board Members :: SBOE Staff :: County Offices :: Search

CHECK YOUR VOTER REGISTRATION HERE

Voter Registration
Voting Information
Data and Statistics
Forms
Election Laws
SEIMS
Related Links

Voter Data Results From The NC Statewide Database

Click Here to Search for Another Voter.

Name:	KARR, ALAN FRANCIS
County Name:	ORANGE
Status:	ACTIVE
City:	CHAPEL HILL NC 27516
Race:	WHITE
Ethnicity:	NOT HISPANIC or NOT LATINO
Gender:	Male
Party:	[illegible]

Finding Date of Birth

AnyBirthday.com

846 West St., New York, NY 10001
Born: Sep. 11, 1902

Smith. John R.

Click here for Addresses and Phone Numbers of your search subject.

Search using Age or birthday

Locateme.com

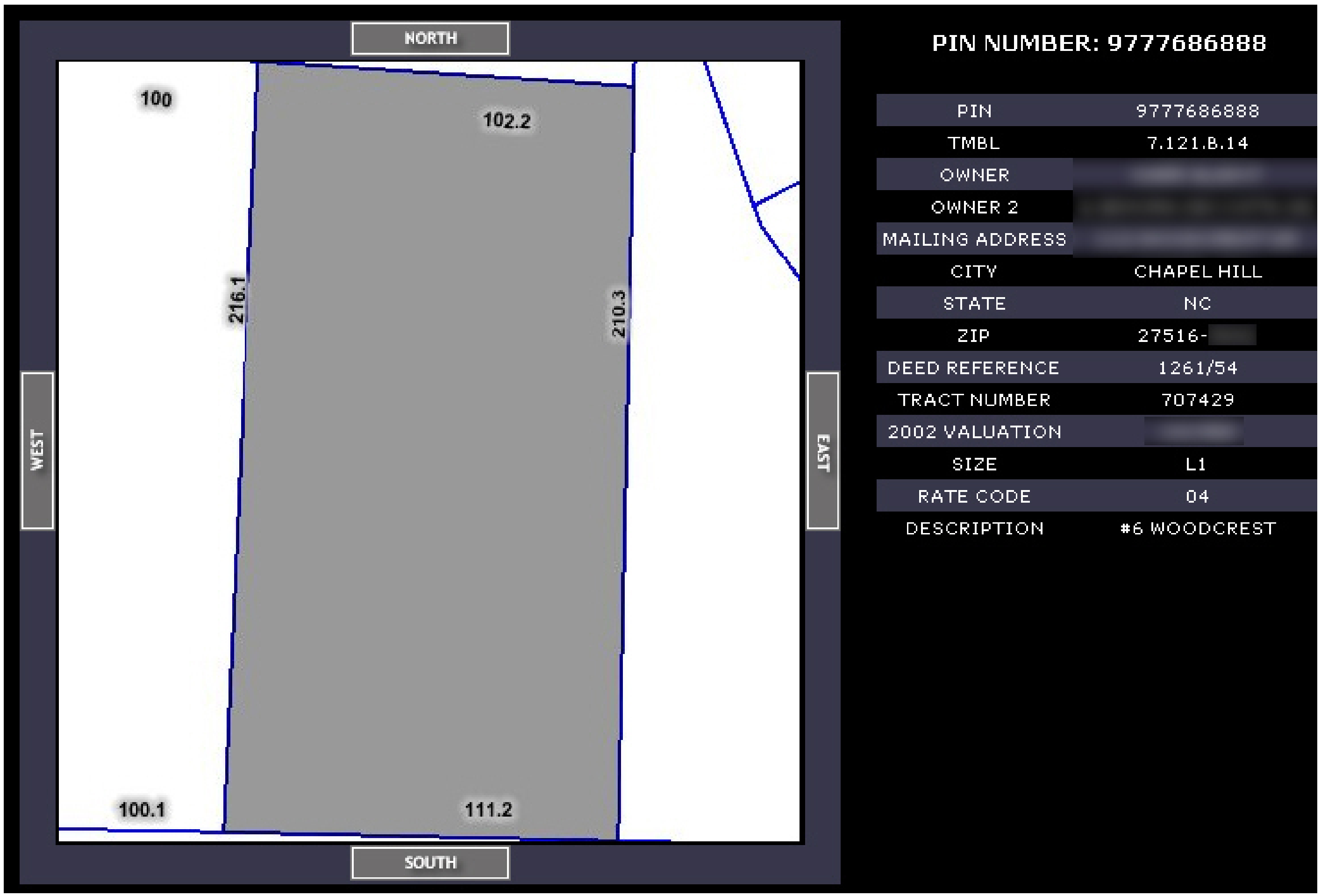
Click here for a Name and Age Search

NEW! **Anybirthday.com PLUS lists Addresses!**

Subject's Name	Birthday	Zip Code
ALAN F KARR	[illegible]	27516

ADDRESS: * Included for *Plus* Users Only Click for Anybirthday *PLUS*

And More …


NORTH
SOUTH
WEST
EAST
100
102.2
216.1
210.3
100.1
111.2
PIN NUMBER: 9777686888
PIN 9777686888
TMBL 7.121.B.14
OWNER [redacted]
OWNER 2 [redacted]
MAILING ADDRESS [redacted]
CITY CHAPEL HILL
STATE NC
ZIP 27516-[redacted]
DEED REFERENCE 1261/54
TRACT NUMBER 707429
2002 VALUATION [redacted]
SIZE L1
RATE CODE 04
DESCRIPTION #6 WOODCREST

Approaches to SDL

- Restricted access
- Restricted data
  - The truth but not the whole truth
- Altered data
  - Nearly the truth
- Analysis servers (DASs)

Restricted Access

- Access at restricted data center [, at a cost]
  - To approved individuals
  - For approved analyses (before and after)
- Advantages
  - Real data
- Disadvantages
  - Inequitable
  - Manual approval processes
    - Costly
    - May ignore important issues such as query interaction

Restricted and Altered Data

- $\mathcal{O}$ = original database
- Restricted data release: $M = f(\mathcal{O})$
  - Often, $M = \{g(r) : r \in \mathcal{O}\}$
  - Examples: drop attribute, coarsen attribute
- Altered data release: $M = f(\mathcal{O}, \text{randomness})$
  - Often, not record-by-record
  - Examples: Microaggregation, data swapping

A Categorization

- Dimension 1: degree of “borrowing”
  - None: M_i depends only on $\mathcal{O}_i$
  - Some: M_i depends on $\mathcal{O}_i$ and small number of other $\mathcal{O}_j$
  - A lot: M_i depends on “all” $\mathcal{O}_j$
- Dimension 2: exogeneous randomness or not
  - Affects data values
  - Inherent in algorithm

Examples of Altered Data

- Aggregation
  - Geographical
  - Coarsening of categories, especially top-coding
- Perturbation
  - Additive (or other) noise
  - Microaggregation
  - Data swapping
- Synthetic data

Risk-Utility Formulations

Generalities

- Components
  - Database $\mathcal{O}$
  - Set $\mathcal{R}$ of candidate releases (masked databases) M
  - Disclosure risk function **DR**(M)
  - Data utility function **DU**(M)
- Goal: Select the “best release”
  - Problem: More utility = more risk

Selection Procedure 1

- Maximize utility subject to upper bound on risk

$$M^* = \arg\max_{M \in \mathcal{R}} \mathbf{DU}(M)$$

$$\text{s.t. } \mathbf{DR}(M) \leq \alpha$$

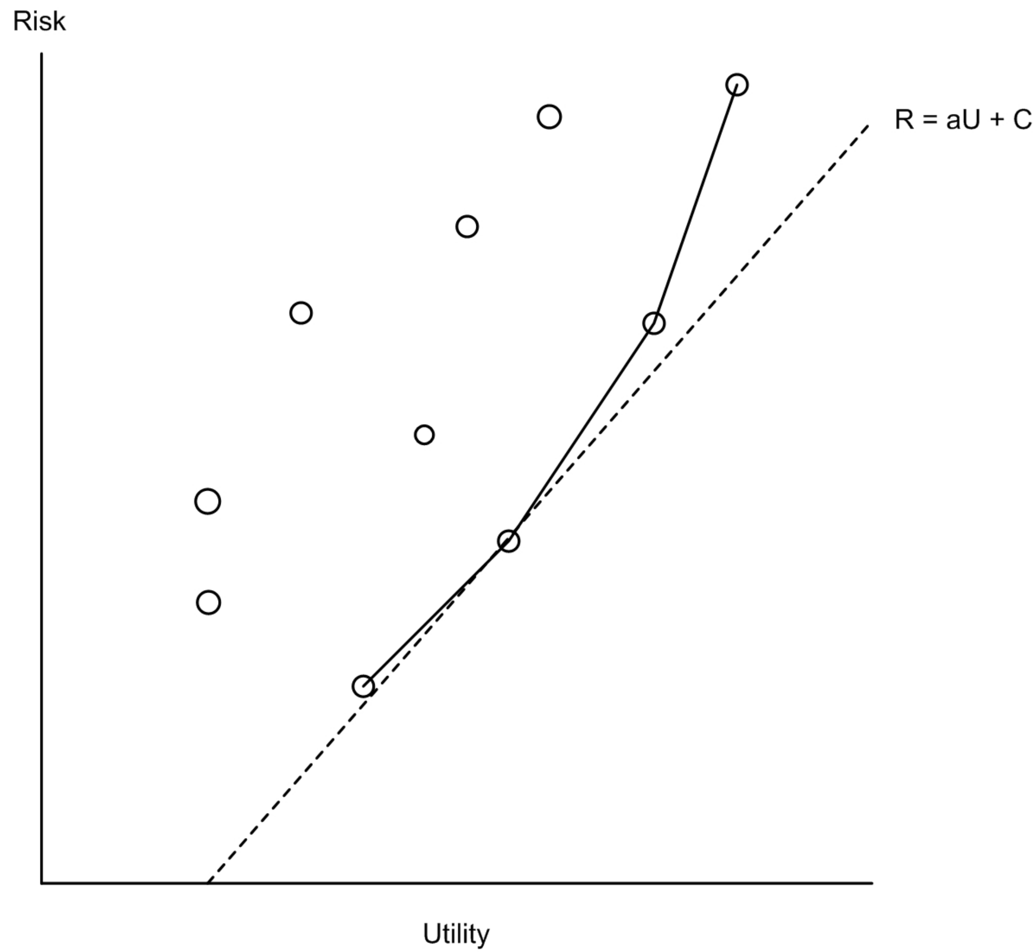
Selection Procedure 2

- Select from *risk-utility frontier* defined by the partial order

$$M_1 \preceq_{\mathrm{RU}} M_2 \quad \Leftrightarrow \quad \mathbf{DR}(M_2) \le \mathbf{DR}(M_1) \text{ and } \mathbf{DU}(M_2) \ge \mathbf{DU}(M_1)$$

- Can use
  - Utility function
  - Other means of choice

Conceptual Risk-Utility Frontier



Risk Measures

- An underlying concept
  - Relative uniqueness = risk
- Tabular data
  - Counts and sums: number of cases in small (e.g., 1 or 2) count cells
  - Sums: dominance (e.g., 60%)
- Microdata
  - Record linkage: % of records linked correctly to parent

Categorical Data: Broad Utility Measures

- Distance between actual tables (Gomatam, et al., JOS, 2005)
  - Hellinger: $\mathsf{HD}(O, M) = \sqrt{\frac{1}{2}\sum_C \left[\sqrt{O(C)} - \sqrt{M(C)}\right]^2}$
  - Total variation
  - Entropy change
- Indistinguishability of *M* from *O* (Woo, et al. JPC, 2006)
  - Propensity scores

Categorical Data: Released Marginals

- Dobra, et al. (IJUFKS, 2002)
  - Number of released marginals
  - Number of released cells
  - Number of degrees of freedom
  - Distance between O and Ô estimated via IPF

Categorical Data: Model-Based

- Likelihoods (Gomatam, et al., JOS 2005)
  - $$\mathbf{DU}_{\mathrm{llm}}(R) = \mathcal{L}_{\mathbf{M}^*}(\mathcal{D}_{\mathrm{post}}(R)) - \mathcal{L}_{\mathbf{M}^*}(\mathcal{D}_{\mathrm{pre}})$$
- Δ(log-linear model) (Denogean, Karr, Qaqish)
- Distance between fitted tables (Denogean, et al., 2006)
  - Model for O → Estimated table Ô
  - Model for M → Estimated table M^
  - Measure distance between Ô and M^

Numerical Data: Broad Measures

- Distance measures
  - Kullback-Liebler (Karr, et al., TAS, 2006)
    - Only feasible (approximately) for normal data
  - Distribution functions (Woo, et al., 2006)
    - Don’t seem to work very well
- Indistinguishability (Woo, et al., 2006)
  - Propensity scores
  - Clustering
  - [SVM]
  - Other classifiers?

Numerical Data: Regression-Specific Measures

- Gomatam, et al., Stat. Sci., 2005
  - Setting: regression servers—protect covariance entries involving X_0 and X_{supp}
  - Dimension of unsuppressed attributes X_{free}
  - $R^2(X_0|X_{\text{free}})$ [+ $\Sigma_{\text{i} \in \text{free}}\, w_i$]
- Karr, et al., TAS, 2006
  - Setting: $X_0|X_{-\{0\}}$
  - Confidence interval overlap
  - Confidence ellipsoid overlap

Example: Geographical Aggregation

- High-resolution geography is a major threat to confidentiality:
  - ZIP code
  - County
  - In some cases, even state
- Possible solutions
  - Report data at regional level (US = 4 regions)
  - Report data at state level
  - Let the data determine the level of aggregation

Data-Dependent Geographical Aggregation

- Annual chemical use survey by National Agricultural Statistics Service (NASS)
  - 194,410 records from 30,500 farms
  - 322 active ingredients
  - 67 crops (field crops, fruits and vegetables)
- Basic reports: application rates (lbs/acre)
- Reporting goal: county level
- Reporting practice: state level

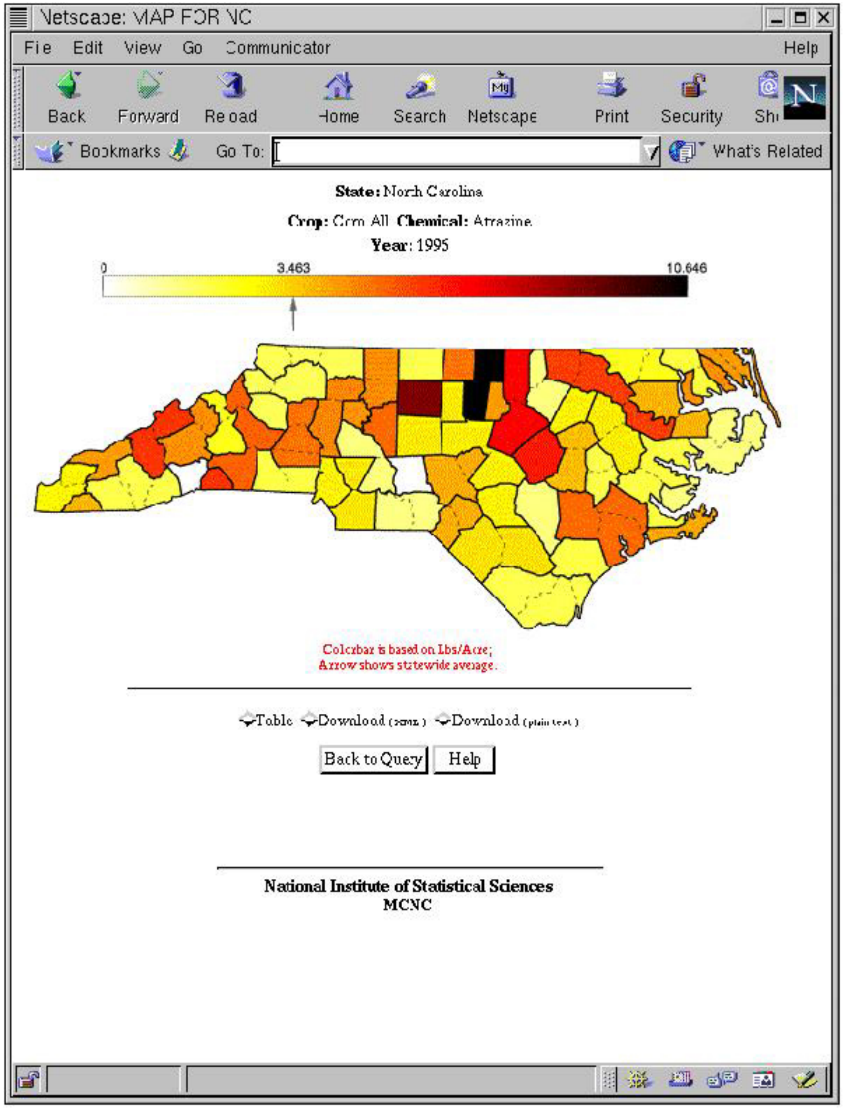
Risk-Utility Formulation

- Release (1 chemical, 1 crop): set *M* of geographical units, and application rate for each
- Disclosure risk: *(N=3, p=.6)* rule. Risk is
  - Infinite if there is a geographical unit
    - Containing fewer than 3 surveyed farms (0 not fewer than 3)
    - In which one farm contains more than 60% of total acreage
  - Zero otherwise

Risk-Utility Formulation

- Data utility: amount of aggregation (less is better)
- Method: aggregate geographically adjacent counties into disclosable "supercounties"
  - "Small" heuristic: minimize size of supercounties
  - "Pure" heuristic: preserve disclosable counties

The Aggregation



Why is Risk-Utility Hard?

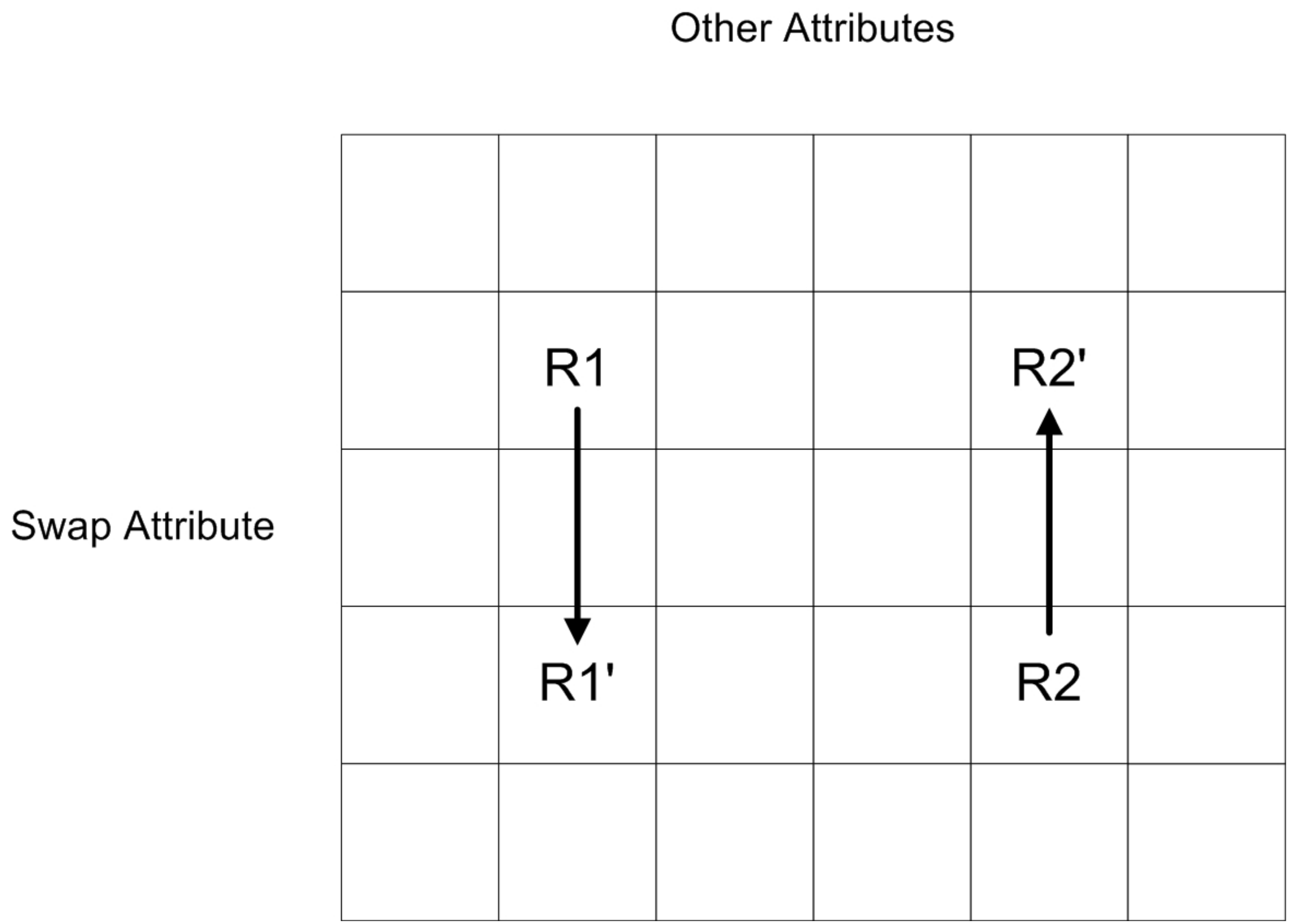
- "One person's risk is another's utility"
  - One difference: incorrect information carries negative utility but positive risk
- Role of external knowledge
  - Risk: disclosure by means of record linkage to external databases
  - Utility: improved analyses by integration with external databases
- Role of data quality

Data Swapping

Data Swapping

- Applied at microdata level
- Basic idea: switch subset of attributes between pairs of records
  - Records, attributes or both can be randomized
- Disclosure risk perspective
  - Reduces risk: intruder cannot be certain that any record is real
- Data utility perspective
  - Distorts data, and so reduces utility

Tabular View



CPS-8: Excerpt from 1993 CPS

- 48,842 data records (not realistic!)
- 8 categorical attributes (not realistic!)
- 2880 cells in full 8-dimensional table (not realistic!)
- 1695 cells with non-zero counts (not realistic!)

Attribute Name (Abbreviation)	Categories
Age (A)	$<$25, 25–55, $>$55
Employer Type (W)	Govt., Priv., Self-Emp., Other
Education (E)	$<$HS, HS, Bach, Bach+, Coll
Marital Status (M)	Married, Other
Race (R)	White, Non-White
Sex (S)	Male, Female
Average Weekly Hours Worked (H)	< 40, 40, > 40
Annual Salary (I)	$<$\$50K, \$50K+

Example Swap for CPS-8

Record	Age	EmplType	Educ	MarStat	Race	Sex	AveHours	Salary
1	<25	Gov	HS	Marr	W	M	40	<$50K
2	25–55	SE	Bach	Marr	NW	M	>40	<$50K
3	25–55	Gov	Bach+	Unmarr	NW	F	>40	>$50K
4	>55	Priv	Bach	Unmarr	W	F	>40	<$50K
5	<25	Other	SomeColl	Marr	W	M	40	>$50K
6	>55	Priv	Bach+	Marr	NW	F	40	>$50K

Record	Age	EmplType	Educ	MarStat	Race	Sex	AveHours	Salary
1	>55	Gov	HS	Marr	W	M	40	<$50K
2	25–55	SE	Bach	Marr	NW	M	>40	<$50K
3	<25	Gov	Bach+	Unmarr	NW	F	>40	>$50K
4	>55	Priv	Bach	Unmarr	W	F	>40	<$50K
5	25-55	Other	SomeColl	Marr	W	M	40	>$50K
6	<25	Priv	Bach+	Marr	NW	F	40	>$50K

Implementation

- Required parameters
  - Swap rate
  - Swap attribute(s)
    - Deterministic (which ones?) or random (probabilities)
  - Record selection randomization probabilities
- Options
  - Constraints on unswapped attributes
  - For numerical data, rank constraints
  - Special treatment for sampling weights
- Domain knowledge checks

Distortion Effects

- For “traditional” (fixed attribute) swapping
  - No change to
    - Joint distribution of swap attributes
    - Joint distribution of unswapped attributes
  - Change to joint distributions that involve both swap and unswapped attributes
- For doubly random swapping
  - All joint distributions change

Risk-Utility Formulation

- Disclosure risk measure

$$\mathbf{DR}(M) = \frac{\sum_{C_1,C_2} \text{Number of unswapped records in } M}{\text{Total number of unswapped records in } M}$$

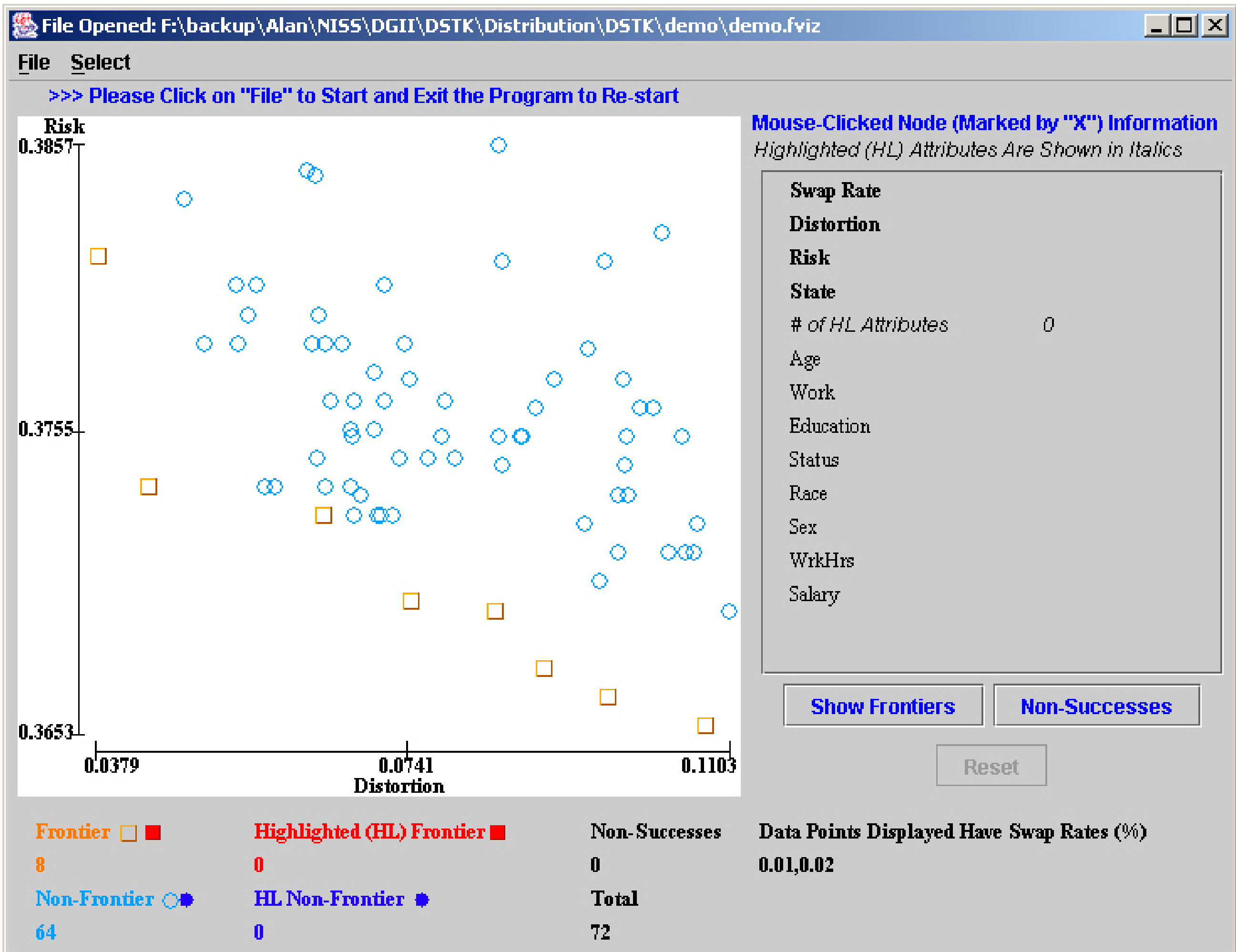
- Utility measure

$$\mathbf{DU}(M) = -\mathbf{DD}(M) = -\mathrm{HD}(\mathcal{D}_{\text{pre}}, M),$$

Data Swapping Experiments

- Done on CPS-8
  - Two rates: 1%, 2%
  - All 8 single-attribute swaps
  - All 28 two-attribute swaps
  - No constraints
- Performed using NISS Data Swapping Toolkit
  - Available at www.niss.org/software/dstk.html

Results


File Opened: F:\backup\Alan\NISS\DGII\DSTK\Distribution\DSTK\demo\demo.fviz
File
Select
>>> Please Click on "File" to Start and Exit the Program to Re-start
Risk
0.3857
0.3755
0.3653
0.0379
0.0741
0.1103
Distortion
Mouse-Clicked Node (Marked by "X") Information
Highlighted (HL) Attributes Are Shown in Italics
Swap Rate
Distortion
Risk
State
# of HL Attributes 0
Age
Work
Education
Status
Race
Sex
WrkHrs
Salary
Show Frontiers
Non-Successes
Reset
Frontier
8
Highlighted (HL) Frontier
0
Non-Successes
0
Data Points Displayed Have Swap Rates (%)
0.01,0.02
Non-Frontier
64
HL Non-Frontier
0
Total
72

Perturbation Methods for Numerical Microdata

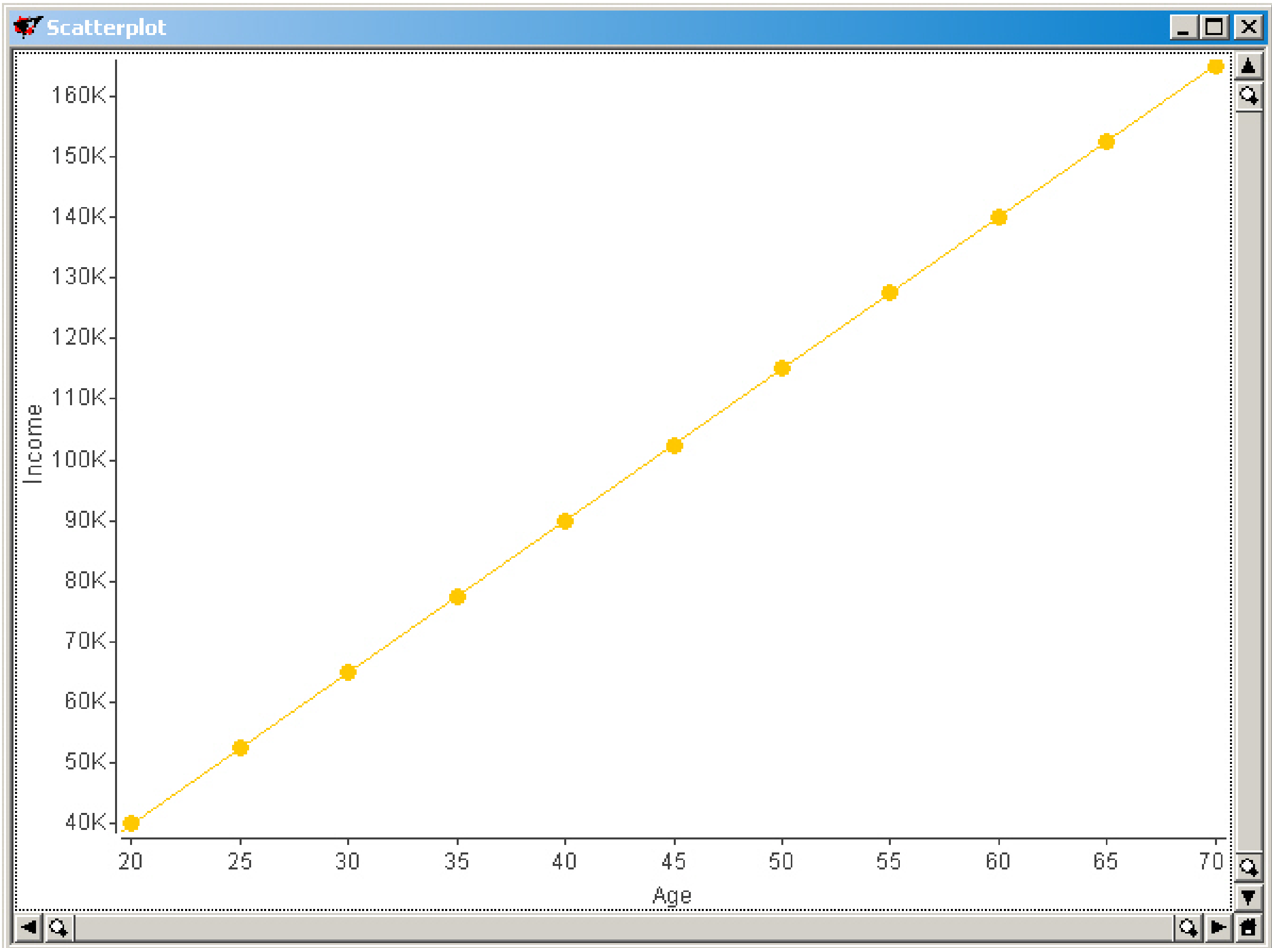
Rationale

- Perturbation can
  - Preserve (robust) statistically interesting low-dimensional relationships in the data
  - Distort (fragile) confidentiality-threatening high-dimensional relationships

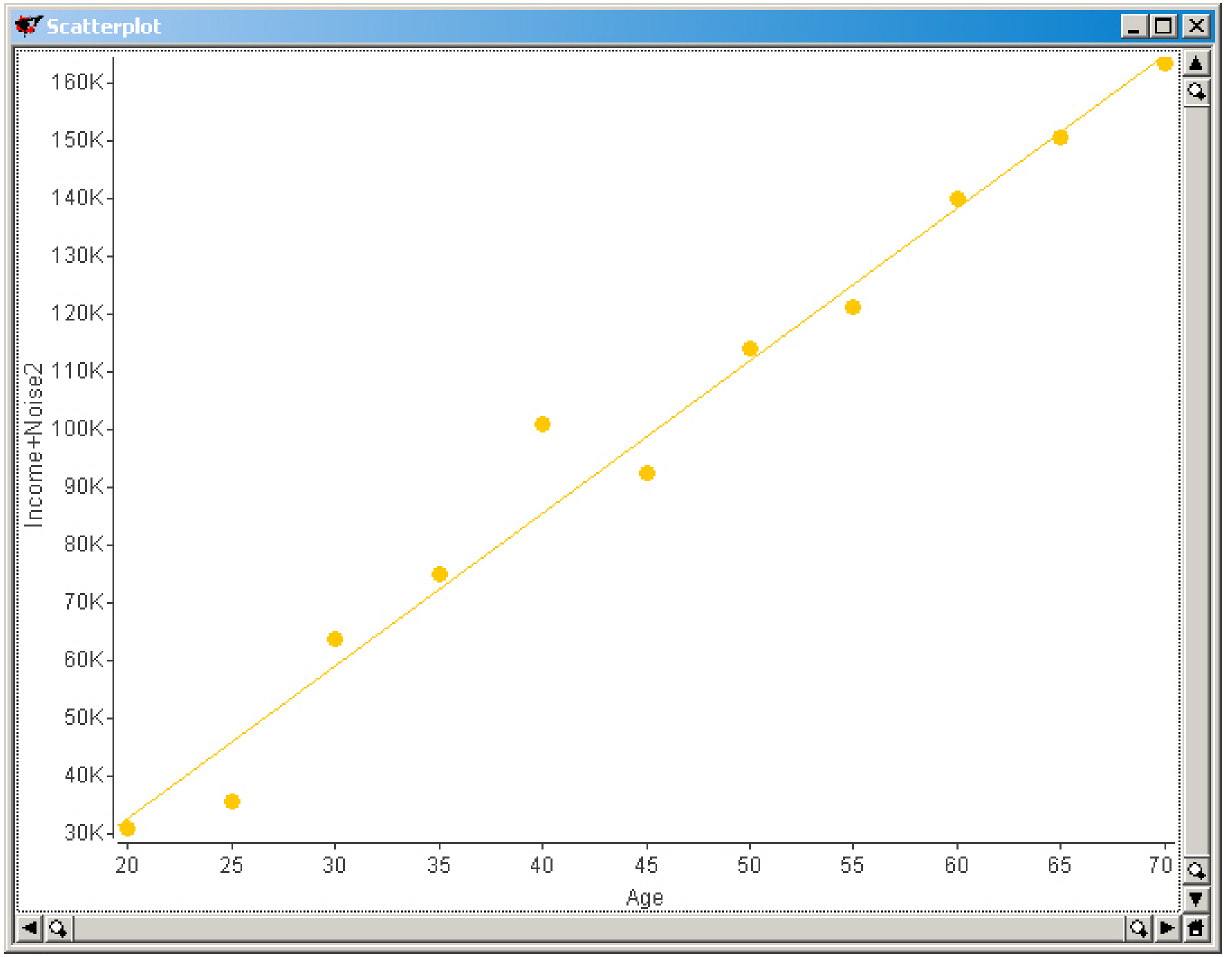
Example: Additive Noise

- Two attributes
  - Age = 20, ..., 70
  - Income
- Assume that age predicts income perfectly
  - Income = $40,000 + $2,500*(Age – 20)
- Add noise to income
  - Uniform on (-5000, 5000)
- Preserves low-dimensional trend
- Destroys high-dimensional exact relationship

Data with No Noise



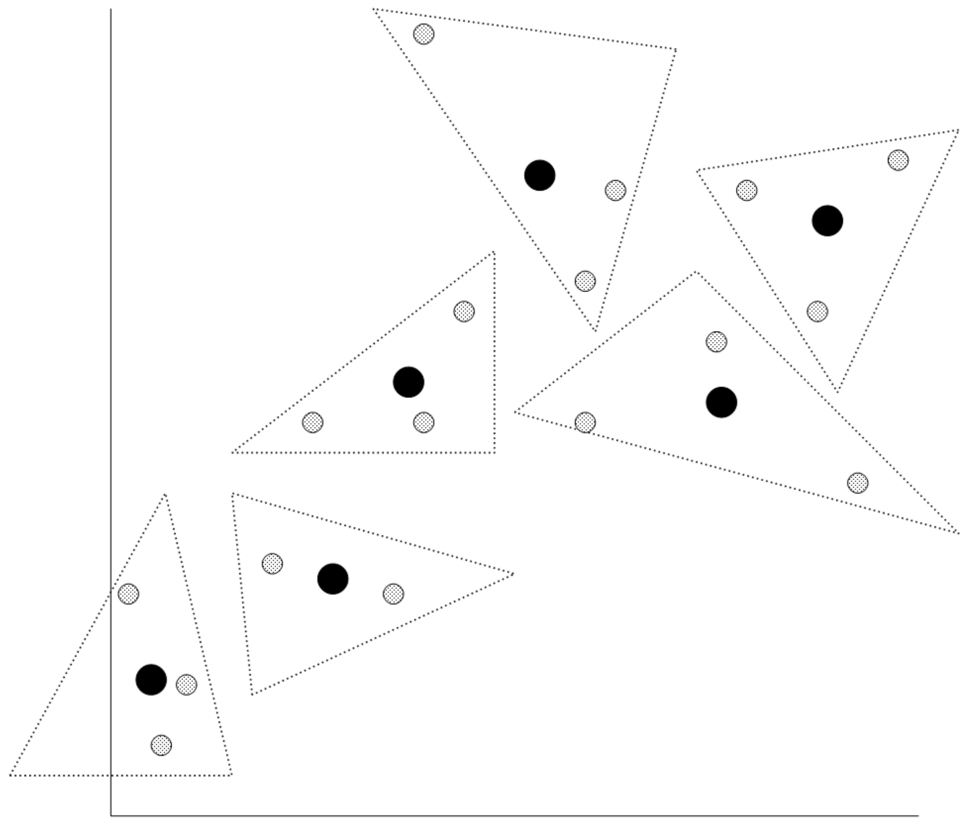
Noise Added to Income



Microaggregation

- Cluster data into sets of size K (K=3 is typical)
- Replace all elements of each cluster by their (attribute-wise) mean
- Choices
  - Clustering algorithm
  - Cluster size

Pictorial Representation



Tabular Data: Background

“Catalog” of Approaches

- Cell suppression
- Release selected marginal sub-tables
- Matrix methods

Example of Cell Suppression

Age / Race	0-20	21-40	40+	Total
White	50	2	25	77
African-American	1	34	15	50
Other	100	101	102	303
Total	151	137	142	430

Original Table (problem cells in red)

Step 1: Suppress Problem Cells

Age / Race	0-20	21-40	40+	Total
White	50	***	25	77
African-American	***	34	15	50
Other	100	101	102	303
Total	151	137	142	430

*** = Primary suppressions

Step 2: Complementary Suppression

Age Race	0-20	21-40	40+	Total
White	####	***	25	77
African-American	***	####	15	50
Other	100	101	102	303
Total	151	137	142	430

= Complementary suppressions

Complementary Suppression Done Dumbly

Age Race	0-20	21-40	40+	Total
White	####	***	####	77
African-American	***	34	15	50
Other	####	101	####	303
Total	151	137	142	430

= Complementary suppressions

What About Marginals Only?

Age Race	0-20	21-40	40+	Total
White				77
African-American				50
Other				303
Total	151	137	142	430

Servers

Servers

- Concept: web-based system to which users submit queries for analyses of $\mathcal{O}$
- Server must
  - Assess risk, taking into account interactions with previously answered queries
  - Assess utility inherent in the query
  - Account for queries that become unanswerable
  - Decide whether (and how) to respond, keeping in mind that a denial may be informative

Fundamentally Different Server Types

- Static: only pre-determined set of queries will be answered
- Dynamic: queries arrive over time, and must be assessed in light of
  - Previously answered queries
  - Queries that would become unanswerable
  - User equity

Server Abstractions

- Database $\mathcal{O}$
- Query space $\mathcal{Q}$
  - Queries $Q = f(\mathcal{O})$ to which the server will respond
- Answer space $\mathcal{A}$
  - If a query Q is not denied, what answer $A(Q)$ is given?
- Released set R ($R(t)$ for dynamic servers)
  - Information contained in all answered queries
- Disclosure risk function **DR**
  - Needs to be defined for *all subsets* of $\mathcal{Q}$!
- Data utility function **DU**
  - Needs to be defined for *all subsets* of $\mathcal{Q}$!

Example: Table Servers

- $\mathcal{O}$ = large (d = 50 dimensions) contingency table
- $\mathcal{Q}$ = all marginal sub-tables of $\mathcal{O}$
- Possible responses to Q
  - A(Q) = refusal to release Q
  - A(Q) = Q, which also releases all sub-tables of Q
- Scalability is a major issue
  - If the table has *d* dimensions, then the number of candidate releases is $\sim 2^{2^d}$

Example of a Static Table Server

- Release R = subset of $\mathcal{Q}$
- Disclosure risk

$$\mathbf{DR}(R) = -\min\{\mathrm{UB}(C, R) - \mathrm{LB}(C, R) : 0 < \#\{C\} \leq 3\},$$

- Data utility

$$\mathbf{DU}(R) = \#\{R\}$$

- Maximize **DU** subject to constraint on **DR**
  - Not solvable because of computational issues: number of releases, calculation of bounds

Example: CPS Data

- 299,285 records
- 13 dimensions
  - 2,592,000 cells
  - 41,672 non-zero cells
  - 22,996 cells with #(C) = 1
  - 6,345 cells with #(C) = 2
  - 3,032 cells with #(C) = 3
- “Optimal” release with width 3 for bounds
  - Frontier = 2 7-way tables and 5 6-way tables
  - Total of 351 sub-tables
- By comparison, release of all 3-way sub-tables contains 377 sub-tables

How Might a Dynamic Table Server Function?

- Queries arrive over time
- Answered query *Q* represents both
  - Direct release: *Q*
  - Possible indirect releases: unreleased children of *Q*
- Totality *R(t)* of released information at *t* described by *released frontier*
- Assume that all users collude
- Need
  - Disclosure risk measure
  - Data utility measure
  - Release rule

More on Dynamic Table Servers

- Maintain disclosure risk below threshold, so refuse to release Q at t if

$$\mathbf{DR}(R(t) \cup Q) > \alpha$$

- Myopic release rule: release Q at t if

$$\mathbf{DR}(R(t) \cup Q) \leq \alpha$$

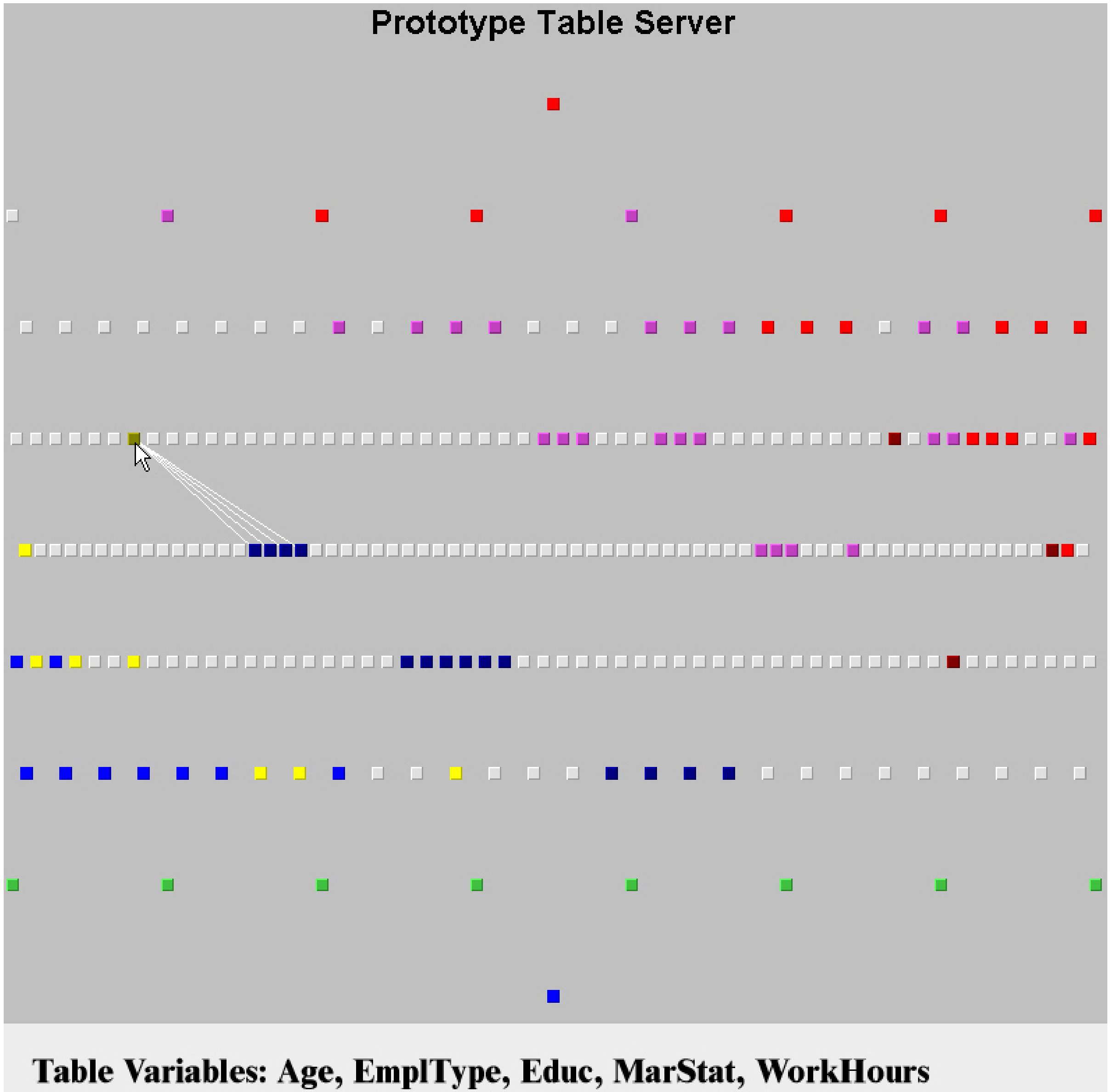
- Could bring in utility by requiring

$$\mathbf{DU}(R(t) \cup Q) - \mathbf{DU}(R(t)) \geq \beta$$

Still More …

- Myopic release rule
  - Fails to account for queries that become permanently unanswerable as a result of answering Q
    - These are specified by an *unreleasable frontier*
  - Cannot prevent small number of users from driving the server into a region that meets their needs
  - Does not naturally accommodate utility
- We don’t know feasible alternatives!

A Pictorial View



Regression Servers

- $\mathcal{O}$ = database of numerical attributes X_i
- $\mathcal{Q}$ = "all regressions" within $\mathcal{O}$:

$$Q = [X_j | X_{i_1}, \ldots, X_{i_k}, i_\ell \neq j]$$

- A(Q) consists of
  - Estimated coefficients
  - Estimated covariance matrix of coefficients
  - R^2
  - ???

Initial Issues

- What about diagnostics?
- Query space $\mathcal{Q}$
  - Is not partially ordered
  - Does not allow transformations of attributes
- Interaction among queries not clear
- What is disclosure risk?
  - Individual data elements
  - Relationships within the data
- What is data utility?

What We Do Know: One Special Case

- X_0 = sensitive variable, $X_1, \ldots, X_d$ = predictors
  - $X_{\mathsf{supp}} \subseteq \{X_1, X_2, \ldots, X_d\}$
  - $X_{\mathsf{free}} = \{X_1, X_2, \ldots, X_d\} \setminus X_{\mathsf{supp}}$
- $\mathcal{Q}$ = all regressions *except* those with
  - X_0 as response
  - Any element of X_{supp} as predictor, and vice versa

$$\mathbf{S} = \left[\begin{array}{c|cc} s_{00} & \mathbf{s}_{\mathrm{supp}}^t & \mathbf{s}_{\mathrm{free}}^t \\ \hline \mathbf{s}_{\mathrm{supp}} & \multicolumn{2}{c}{\multirow{2}{*}{$\mathbf{S}_D$}} \\ \mathbf{s}_{\mathrm{free}} & & \end{array}\right]$$

Disclosure Risk

- Residual risk $\mathbf{DR}_{\text{res}}$
  - 1 / square root of the average of the squared residuals for selected subset (e.g., those with extreme attribute values) of the data
- Prediction risk $\mathbf{DR}_{\text{pred}}$
  - Draw feasible values of $\mathbf{s}_{\text{supp}}$ from the ellipsoid to which they are constrained by X_{free}, generating feasible coefficients for $[X_0|X_1,\ldots,X_d]$
  - Risk measure is the average value of R^2 for these feasible regressions

Data Utility

- Unweighted utility
  - $\mathbf{DU}_{\mathrm{rsq}} = R^2$ of $[X_0|X_{\mathrm{free}}]$
- Weighted utility
  - $\mathbf{DU}_{\mathrm{rsqwt}} = U_{\mathrm{rsq}} + \Sigma_{\mathrm{free}}\, w_i$
  - Way of incorporating domain knowledge
- Problem: risk and utility are hard to differentiate

Research Frontiers

Playbill

- New measures of utility and risk
- Combining SDL methods
- [Transparency]
- Distributed databases

New Utility Measures

- The basic idea
  - Merge original data $\mathcal{O}$ and masked data M, each labeled
  - Attempt to classify them without using the labels
  - If not successful, then M is a good surrogate for $\mathcal{O}$
- Classification methods
  - K-means: So-so
  - Distribution functions: not very good
- Propensity scores: really good
  - Scalability issue: involves model

Axioms for Data Utility

- Basis: Consumer preference theory in microeconomics
- ***DU*** = data utility measure
  - ***DU****(M)* = utility of masked data *M*
- **Anchoring**
  - ***DU****(Ø) = 0*
- **Satiation**
  - ***DU****(𝒪)* ≥ ***DU****(M)* for all *M*
- **Usefulness of *𝒪***
  - Easy: ***DU****(𝒪) > 0*
  - Admissibility: *M* such that ***DU****(M)* ≥ ***DU****(Ø)*

More Candidate Axioms

- **Monotonicity**
  - In $n(M)$: if M' contains a subset of the rows of M but the same columns, then $\mathbf{DU}(M) \geq \mathbf{DU}(M')$
  - In $k(M)$: if M' contains a subset of the columns of M but the same rows, then $\mathbf{DU}(M) \geq \mathbf{DU}(M')$
  - In parameters $p($SDL method$)$ with the “form” of the method fixed
- **Convexity**, as a stronger form of monotonicity

What are the Real Questions?

- Broad (= blunt) vs. specific (= narrow)
- Is there anything meaningful in-between?
- Scalability
  - Dimension
  - Data set size
- How does domain knowledge enter?
  - Example: Variable transformations

More Questions

- Utility implications of transparency
  - Altered analyses
  - Proper accounting for SDL-induced uncertainty
- What principles should be used to choose utility measures?
  - Is utility one-dimensional?
- Entirely new approaches to utility?
  - Example: tied to decisions based on data, not data per se

Distribution Function Measures of Risk

- The idea: risk of a masked data set M is measured by a distribution function $F_M(t)$
- Can compare candidate releases using stochastic ordering

$$M_1 \preceq M_2 \text{ if } F_{M_1}(t) \leq F_{M_2}(t) \text{ for all } t$$

- Can define frontier

Example 1

- $r =$ record-level measure of risk (e.g., probability of re-identification)
- $w(x) =$ importance of record x, scaled so that $w(x) \geq 0$ and $\sum_x w(x) = 1$
- $F_M(t) = \sum_{x \in M} w(x) 1\{r(x) \leq t\}$

Example 2

- K = intruder's knowledge = unobserved random variable
- P = agency prior on K
- $r(M|K)$ = risk of masked data M when intruder's knowledge is K
- $F_M(t) = \int 1\{r(M|k) \leq t\} dP(k)$

Example 3

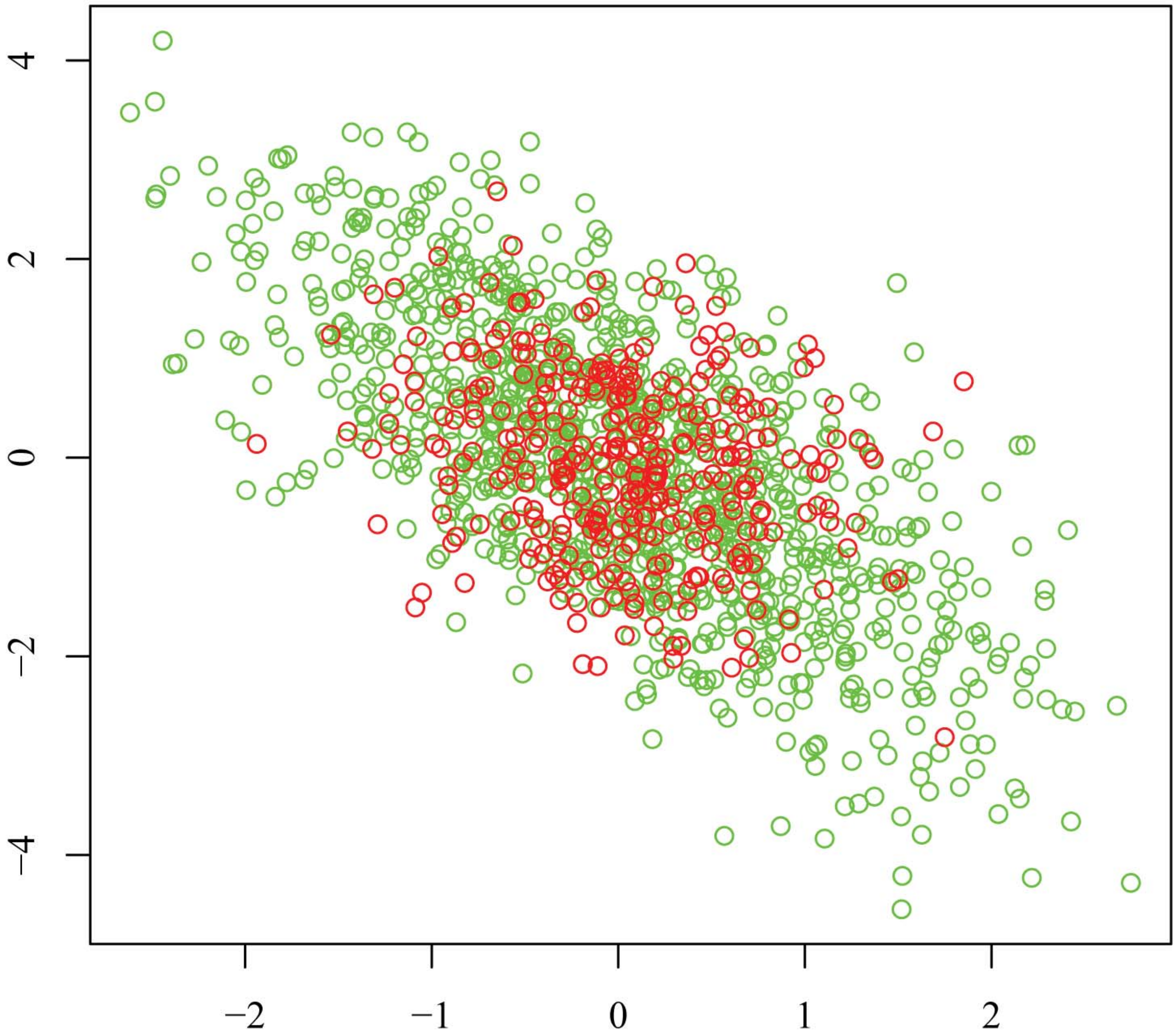
- Multiple risk measures $R_1(\cdot), \ldots, R_J(\cdot)$
- Weights $w_1, \ldots, w_J$
- Consider $F_M = \sum_j w_j \varepsilon_{R_j(M)}$
- Many measures are of the form $\int f(t) dF_M(t)$
- Example: $f(t) = t \rightarrow R(M) = \sum w_j R_j(M)$

Combining Methods

- First, apply a method
  - That is good for risk
  - Whose utility consequences can be characterized
- Second, apply a method that
  - Undoes what the first method did to utility
  - Does not undo what the first method did for risk

Example

- Original data $\mathcal{O}$
- Stage 1: microaggregation to produce M_1
  - Good for risk
  - Reduces covariance
- Stage 2: additive noise to restore lost covariance, which can be done "intelligently"
  - Example: $M_2 = M_1 + N$, where $\mathrm{Cov}(N) = \mathrm{Cov}(\mathcal{O}\text{-}M_1)$


4
2
0
−2
−4
−2
−1
0
1
2

Simulation Study

- 8-variable numerical databases, with varying correlation structures
- Utility: propensity score
- Risk: record linkage
- Masking method
  - Stage 1: microaggregation with *z*-scores projection
  - Stage 2: multiple methods
    - Microaggregation with *z*-scores projection
    - Microaggregation with principal components projection
    - Multivariate microaggregation
    - Rank swapping
    - Noise

Propensity Score Utility

	Symmetric				Non-symmetric			
	High Corr		Low Corr		High Corr		Low Corr	
	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos
MicZ	281.51	128.14	233.40	132.12	639.42	592.07	639.04	463.78
MicZ-Noise	26.49	9.00	16.97	92.7	15.5	5.69	7.53	28.75
MicZ-MicMul	16.53	18.37	12.97	14.84	91.5	5.48	8.68	11.15
MicZ-MicPCP	9.31	12.83	9.33	7.86	3.39	4.99	5.76	8.61
MicZ-MicZ	28.30	2348	33.92	37.27	180.10	45.09	94.10	40.04
MicZ-Rank	34.81	28.58	29.42	27.26	42.04	14.82	26.45	39.48

Disclosure Risk

	Symmetric				Non-symmetric			
	High Corr		Low Corr		High Corr		Low Corr	
	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos
MicZ	.0025	.0036	.0019	.0024	.0011	.0043	.0011	.0012
MicZ-Noise	.0044	.0077	.0044	.0076	.0079	.0144	.0098	.0039
MicZ-MicMul	.0046	.0077	.0025	.0203	.0947	.1122	.1265	.0071
MicZ-MicPCP	.2516	.0198	.0029	.9133	.2926	.0806	.1800	.0477
MicZ-MicZ	.0015	.0275	.0023	.0035	.0004	.0033	.0009	.0477
MicZ-Rank	.0119	.0092	.0067	.0087	.0079	.0340	.0091	.0096

More General Approach

- Attempt to solve

$$f_{\mathcal{O}} = g * f_{M_1}$$

- Release $M_1 + N$ where $N \sim g$
- Can be done numerically, but only in low dimensions

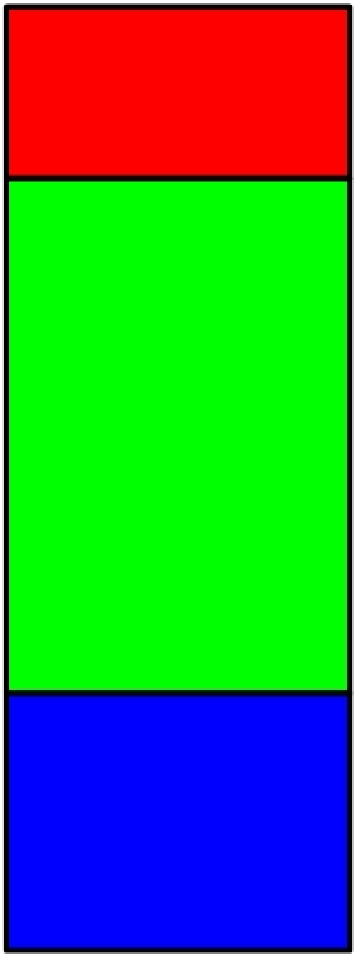
Very General Approach

- $M_1 =$ masked version of $\mathcal{O}$
- $M_2 =$ masked version of $\mathcal{O} - M_1$
- $\ldots$
- $M_k =$ masked version of $\mathcal{O} - M_1 - \ldots - M_{k-1}$
- Release $M = \sum_{i=1}^{k} M_i$

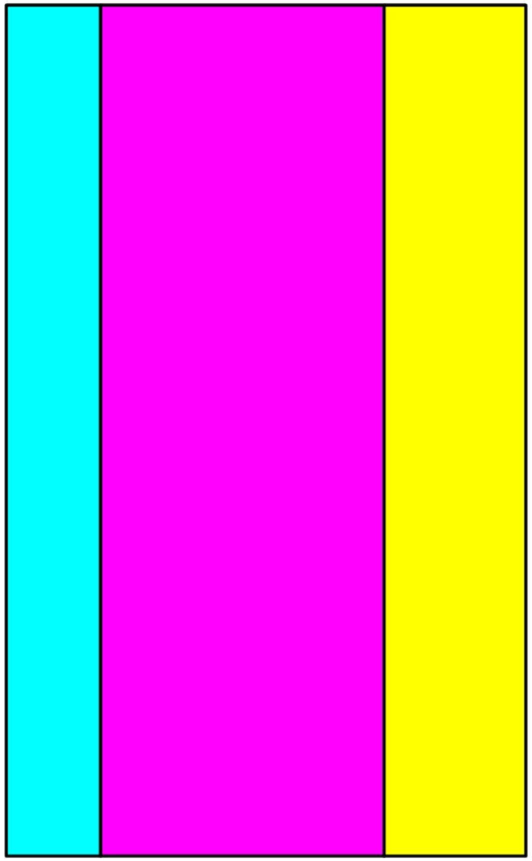
Distributed Data: Problem Formulation

- Multiple, distributed [, related] databases held by different “owners”
  - Government agencies (example: US states)
  - Corporations (example: pharmaceutical companies)
- Goals
  - *Valid statistical inference* on the “integrated” database without actually creating it
  - Protect each owner’s data from the other owners
  - [Protect data subjects]
- Constraints
  - No trusted third party (human or machine)
  - Semi-honest owners

Data Partitioning Models



Horizontal



Vertical

Semi-Honesty

- Database owners
  - *Must* use correct data
  - *Must* perform agreed-on computations properly
  - *May* retain results of intermediate computations

Secure Summation

- Problem
  - Agency k has v_k
  - Agencies want to compute Σv_k in such a way that All agency j learns about other agencies' values is what can be deduced from v_j and the global sum
- Solution
  - Agency 1: generate enormous random number R, and transmit $R + v_1$ to agency 2
  - Agency 2: Add v_2, transmit $R + v_1 + v_2$ to agency 3
  - ...
  - Agency 1: receive $R + \Sigma v_k$, subtract R, share result

Simple Application: Secure Average

- Each agency has income data, and they want to calculate the global average income
  - n_j = number of subjects for agency j
  - I_j = total income for agency j
- Use secure summation to compute and share
  - $I = I_1 + \ldots + I_K$
  - $n = n_1 + \ldots + n_K$
- Each agency then computes I/n

Weaknesses of Secure Summation

- Needs "good" random number
- Collusion is possible
  - Agencies *n-1* and *n+1* can share information and determine a_n without revealing a_{n-1} or a_{n+1}
  - Can be defeated by
    - Splitting calculation into pieces, with different orders for each
    - Hiding order from agencies, as in NISS SCS
- Breaks if semi-honesty fails
  - More later

Regression for Horizontally Partitioned Data

- Setting: Agencies hold same numerical attributes on disjoint sets of subjects
  - y = response
  - X = predictors
- Goal: Fit the linear regression $y = X\beta + \varepsilon$ *including diagnostics*
- Constraints
  - As above

Solution via Secure Summation

- Compute

$$X^T X = \sum_{j=1}^{K} (X^j)^T X^j \qquad X^T y = \sum_{j=1}^{K} (X^j)^T y^j$$

entrywise by secure summation

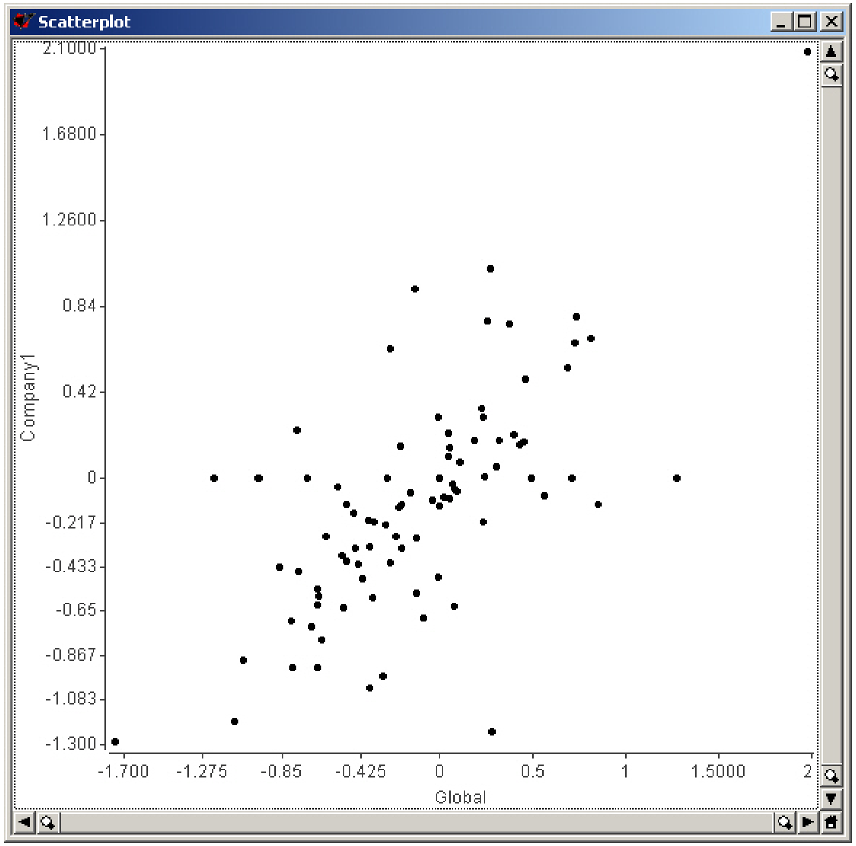
- Share these among agencies; each calculates

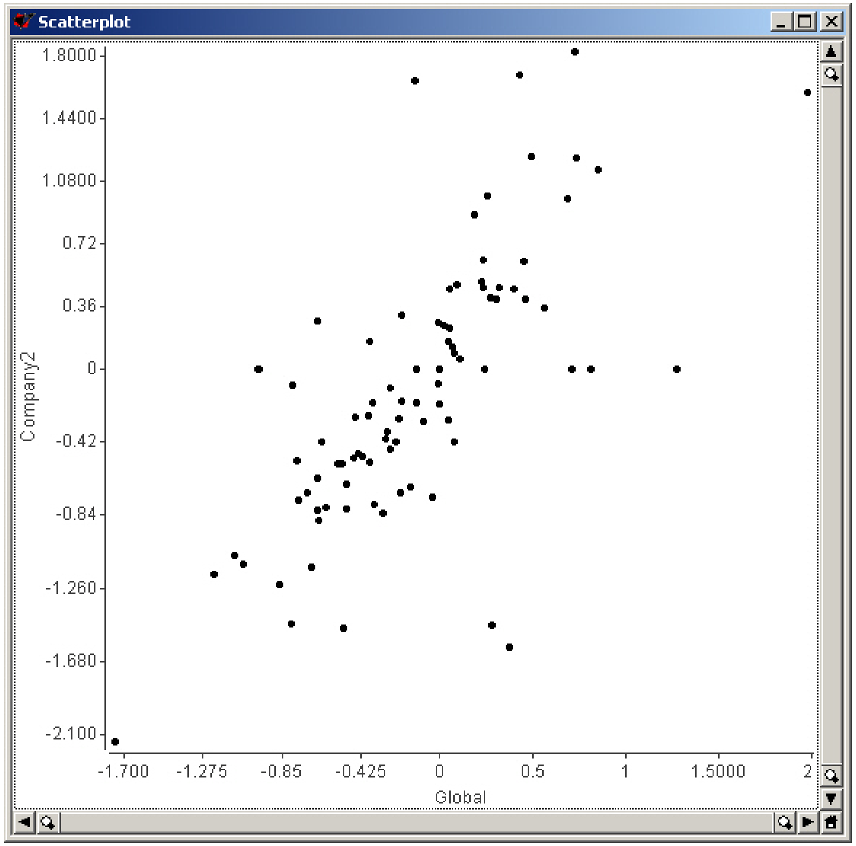
$$\widehat{\beta} = (X^T X)^{-1} X^T y$$

Example: Chemical Data from Multiple Pharmaceutical Manufacturers

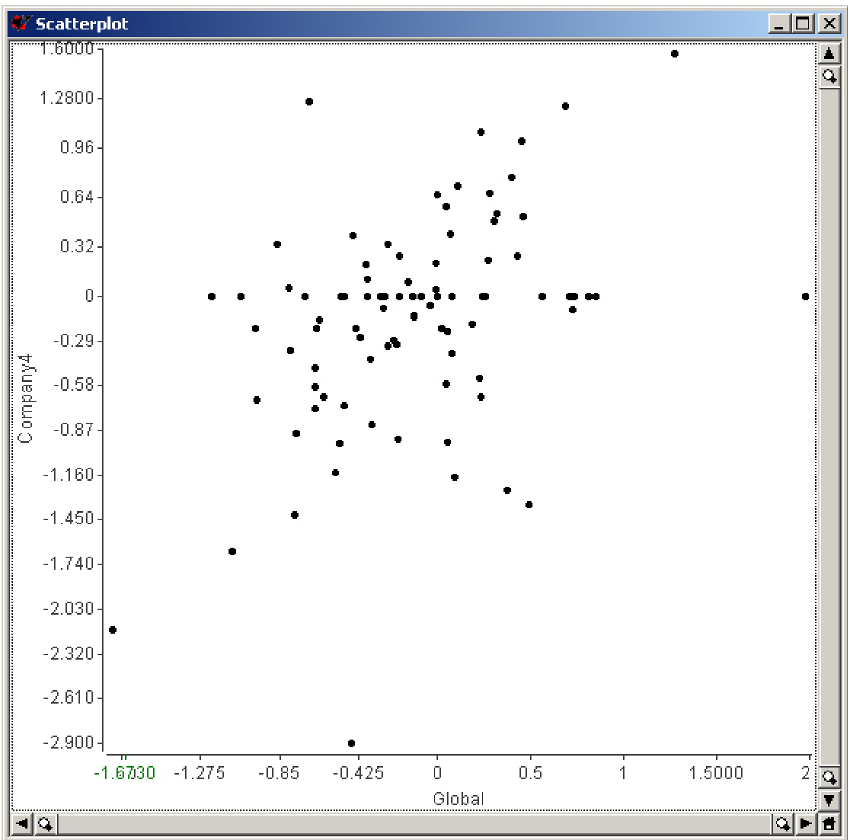
- Data
  - 1318 molecules
  - Response: water solubility
  - Predictors
    - 1 constant
    - 90 (binary) molecular descriptors
- 4 “synthesized” companies
  - Data split using classifier, so each company’s data are relatively homogeneous, but with gaps!
  - Numbers of molecules = 499, 572, 16 (!), 231

Results

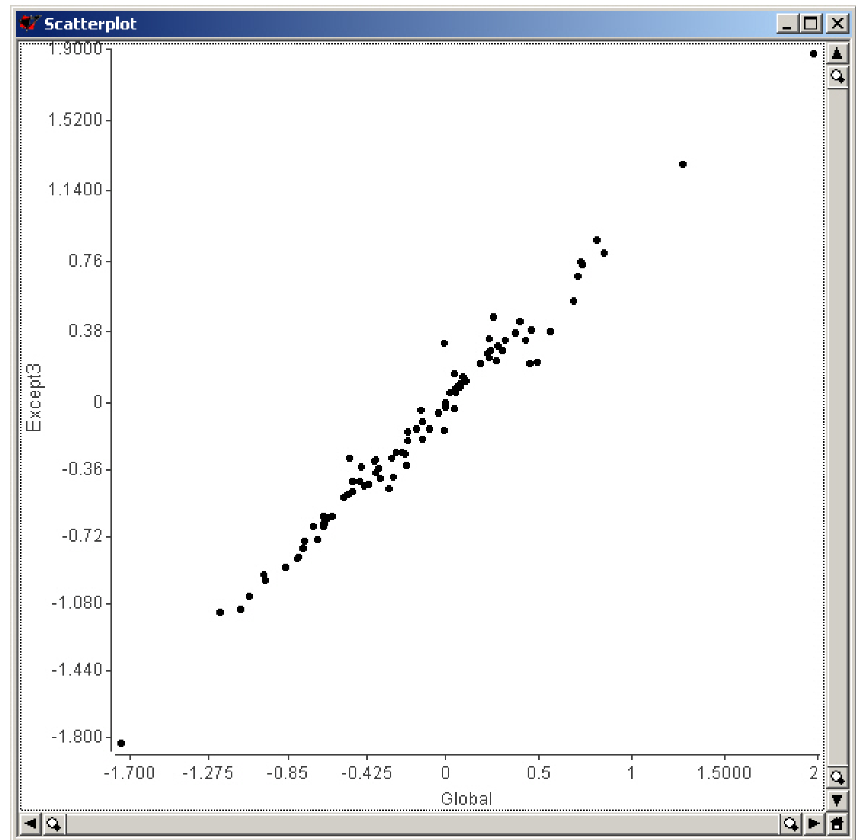

Scatterplot
2.1000
1.6800
1.2600
0.84
0.42
0
-0.217
-0.433
-0.65
-0.867
-1.083
-1.300
Company1
-1.700
-1.275
-0.85
-0.425
0
0.5
1
1.5000
2
Global


Scatterplot
1.8000
1.4400
1.0800
0.72
0.36
0
-0.42
-0.84
-1.260
-1.680
-2.100
Company2
-1.700
-1.275
-0.85
-0.425
0
0.5
1
1.5000
2
Global

Results—2

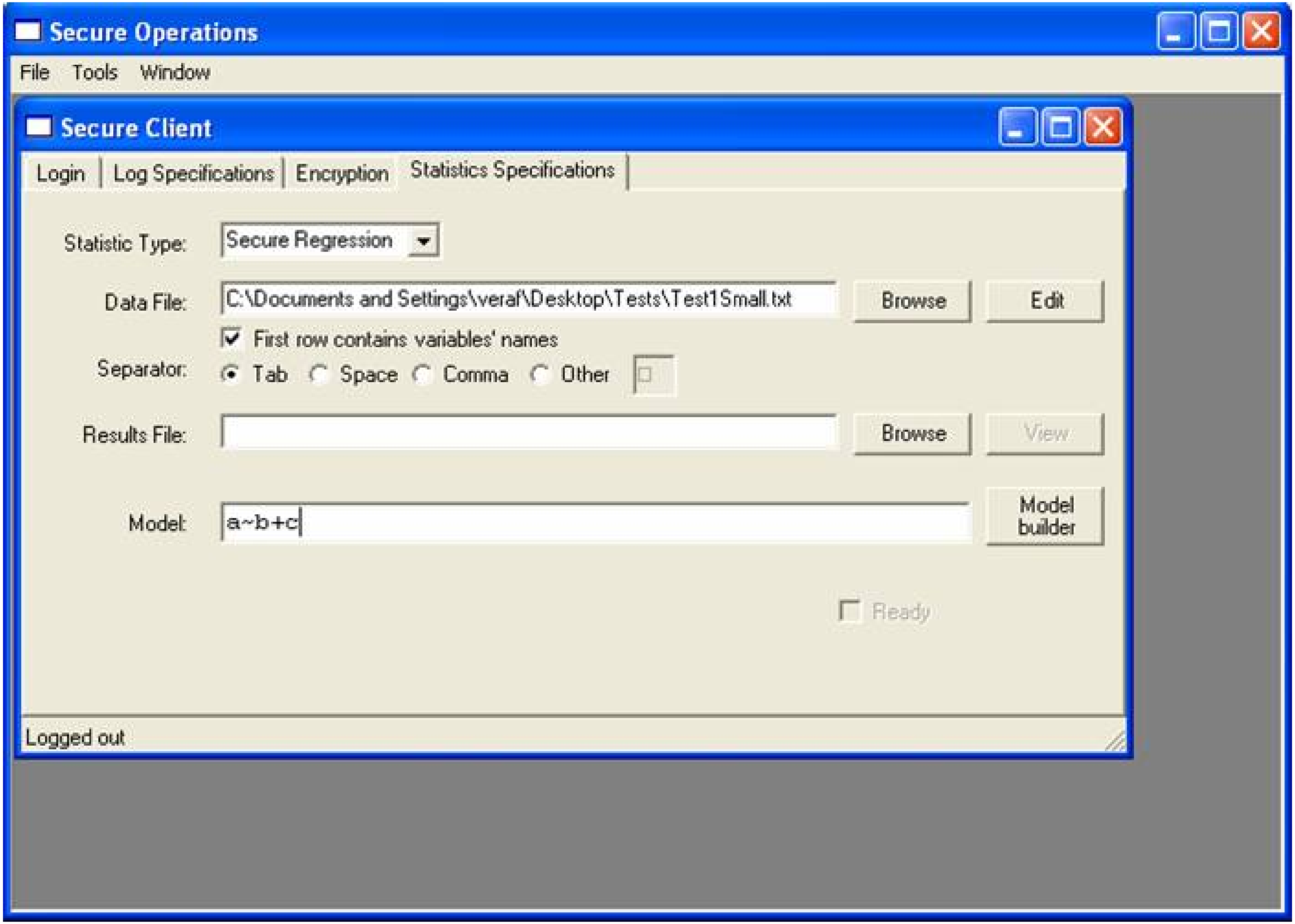


Scatterplot
1.6000
1.2800
0.96
0.64
0.32
0
-0.29
-0.58
-0.87
-1.160
-1.450
-1.740
-2.030
-2.320
-2.610
-2.900
Company4
-1.275
-0.85
-0.425
0
0.5
1
1.5000
2
Global



Scatterplot
1.9000
1.5200
1.1400
0.76
0.38
0
-0.36
-0.72
-1.080
-1.440
-1.800
Except3
-1.700
-1.275
-0.85
-0.425
0
0.5
1
1.5000
2
Global

NISS Secure Computation System



SCS: Regression Output

RegressionViewer

Formula: 1+2+3~4+5+6+7+8+9+10+11+12+13+14+15

Model builder

Analysis for response: 1

1
2
3

Analysis of Variance Table

Source of Variation	Degrees of Freedom	Sum of Squares	Mean of Squares	F-Statistic	p-value
Model	12	1067.0529	88.9211	96.7261	<0.0001
Error	1305	1199.6975	0.9193		
Total	1317	2266.7504			

Coefficients Table

Term	Coefficients	Standard Error	t-statistic	p-value
(Intercept)	0.6064	0.0319	19.0217	<0.0001
4	-0.0129	0.1808	-0.0711	0.9433
5	0.0327	0.0330	0.9907	0.3220
6	-0.1629	0.1150	-1.4175	0.1566
7	-0.0519	0.4408	-0.1178	0.9062
8	0.6995	0.0868	8.0547	<0.0001
9	0.5525	0.0925	5.9722	<0.0001
10	1.0771	0.1776	6.0636	<0.0001
11	2.3014	0.1698	13.5522	<0.0001
12	-0.1835	0.0463	-3.9604	<0.0001
13	0.1733	0.1268	1.3673	0.1718
14	-0.0894	0.1248	-0.7165	0.4738
15	-0.2138	0.0767	-2.7867	0.0054

Thanks to

- Funders of the research: BLS, BTS, Census, NASS, NCES, NCHS, NSF
- Principal collaborators: George Duncan (CMU), Stephen Fienberg (CMU), Sallie Keller-McNulty (LANL), Michael Larsen (ISU), Jerome Reiter (Duke), Lynne Stokes (SMU), Stanley Young (NISS)
- NISS postdocs: Max Buot, Adrian Dobra, Shanti Gomatam, Jaeyong Lee, Xiaodong Lin, Anna Oganian, Ashish Sanil, Francisco Vera, Mi-Ja Woo
- NISS student interns: Jimmy Fulp, Chunhua Liu

References (Easy Way)

www.niss.org/dgii/techreports.html