Secure Statistical Analysis of Distributed Databases (Emphasizing What We Don't Know)

Alan F. Karr
National Institute of Statistical Sciences
Research Triangle Park, NC
karr@niss.org

Data Confidentiality: The Next Five Years
May 1, 2008

NISS
National Institute of Statistical Sciences

Outline

NISS
National Institute of Statistical Sciences

Problem Formulation

Context Related, distributed databases held by multiple owners

- Government agencies (example: US states)
- Corporations (example: pharma companies)

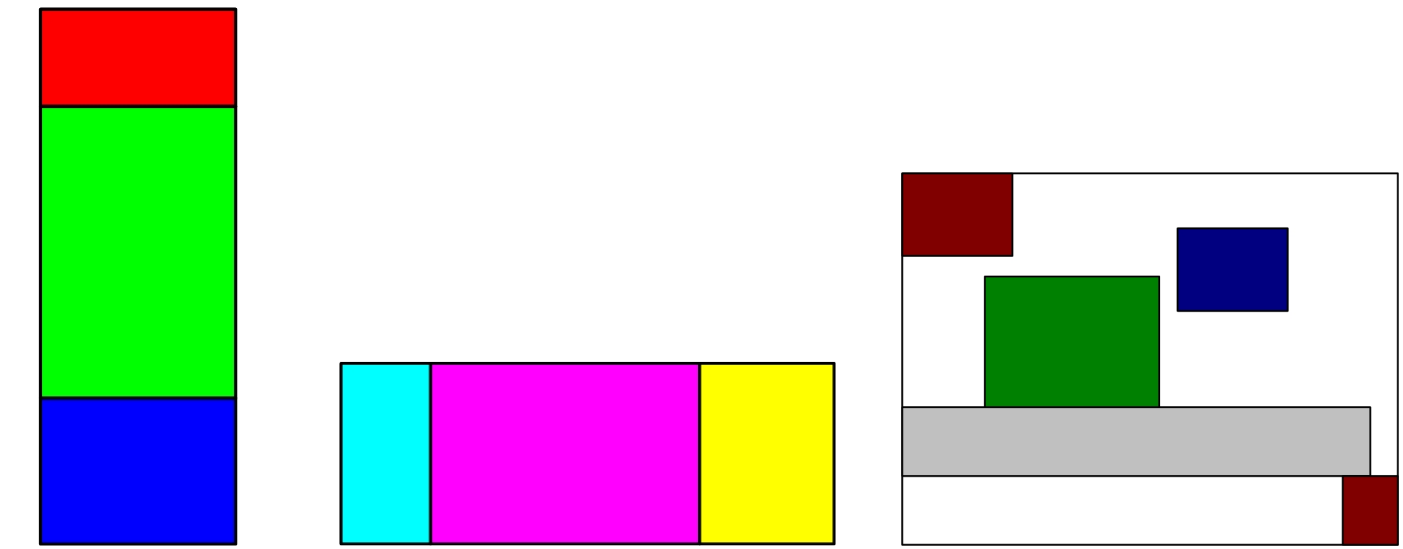
Goal Valid, complete, statistical inference on the integrated database without actually creating it

Constraints

- No trusted third party (human or machine)
- Protect each owner's data from the other owners
- Protect data subjects

Other Considerations Reduce incentives to cheat

Data Partitioning



Secure Multi-Party Computation

Setting
- Agencies $1, \ldots, K$ with values $v_1, \ldots, v_K$
- Known function f with K arguments

Goal Compute $f(v_1, \ldots, v_K)$ (exactly) in such a way that
- All agency j learns about $\{v_i : i \neq j\}$ is what can be deduced from v_j and $f(v_1, \ldots, v_K)$
- Outside parties are not involved

Semi-honesty Agencies
- *Must* use correct data
- *Must* perform agreed-on computations
- *May* retain results of intermediate computations

Secure Summation Protocol

Problem Agencies want to compute $v = \sum_{k=1}^{K} v_k$

Secure Summation Protocol

- *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
- $\vdots$
- *Agency K:* Receive $R + v_1 + \cdots + v_{K-1}$ from agency $K - 1$, add v_K, transmit $R + v_1 + \cdots + v_K = R + v$ to agency 1
- *Agency 1:* receive $R + v$, subtract R, share result

Issues with Secure Summation

- Needs “good” random number
- Collusion
  - Agencies $j-1$ and $j+1$, without sharing private information, can determine v_j
  - Can be defeated by splitting calculation into pieces, with different orders for each
- “Bullet-proof” implementation is subtle
- Breaks if semi-honesty fails: secure summation is not a Nash equilibrium

The One Idea

If the analysis uses sufficient statistics that are additive across agencies, then

1. Use secure summation to compute and share each sufficient statistic
2. Each agency completes the analysis on its own

Linear Regression

Setting Numerical attributes: y = response, X = predictors

Goal Fit the linear regression and $y = X\beta + \varepsilon$ and calculate estimators, diagnostics, ...

Computation of $\hat{\beta}$ via secure summation: Compute

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

and

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

entrywise. Each agency then calculates $\hat{\beta} = (X^T X)^{-1} X^T y$

Example: Chemical Data from Multiple Pharmaceutical Manufacturers

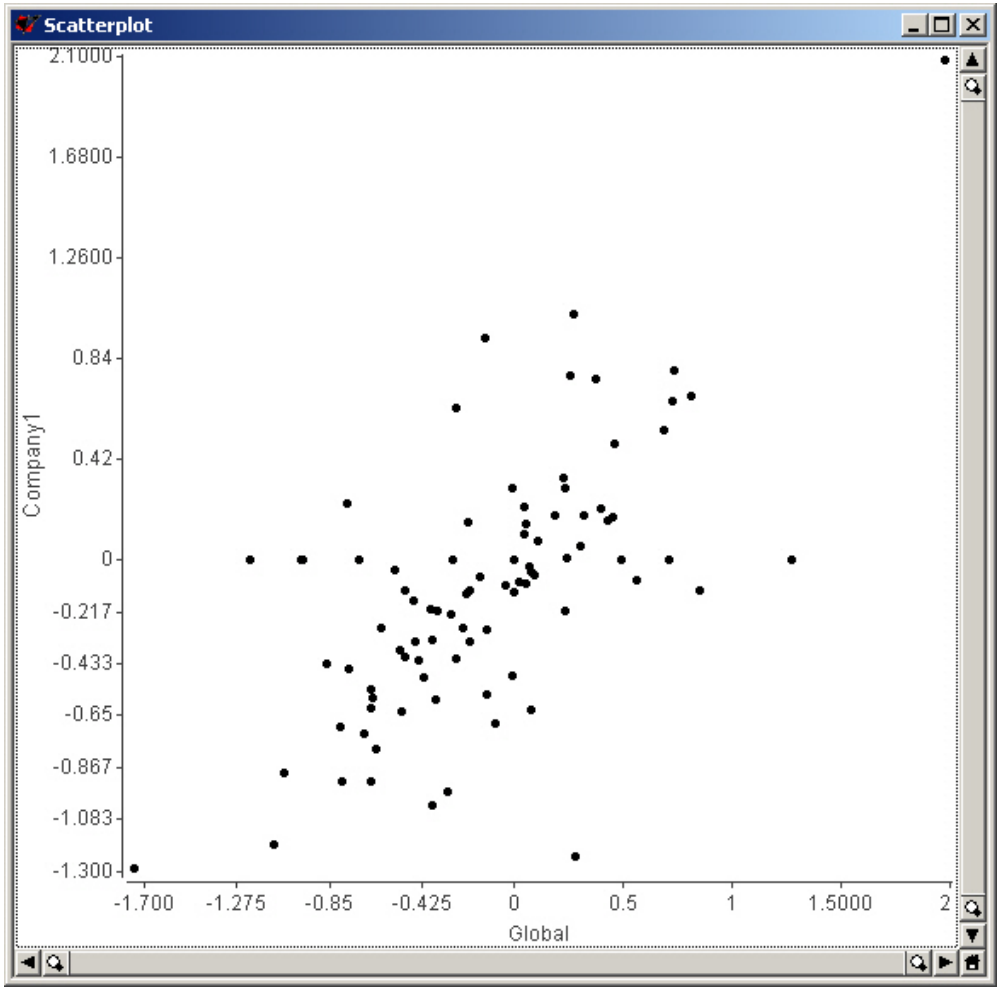
Data 1318 molecules

- Response: water solubility
- Predictors 1 constant + 90 (binary) molecular descriptors

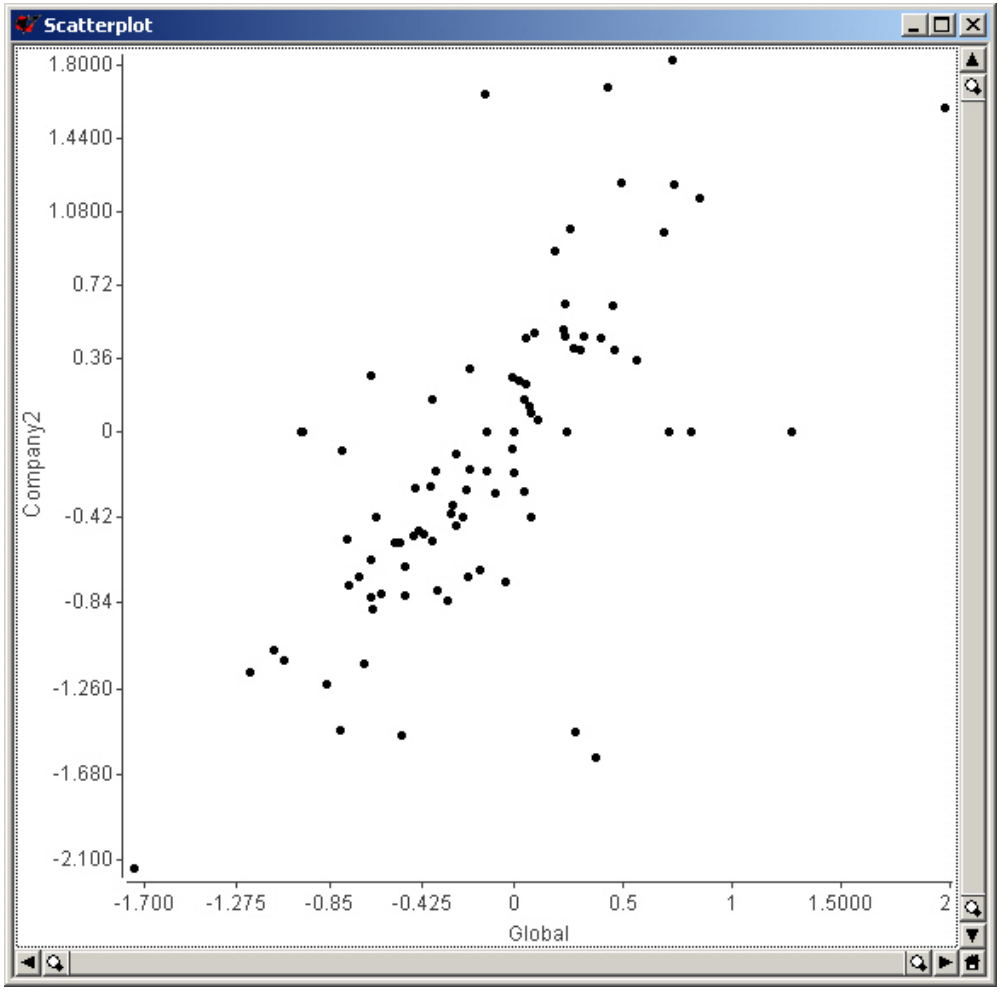
Distributed databases 4 companies

- Data split using classifier: each company's data are homogeneous, but with gaps!
- Sizes of databases = 499, 572, 16 (!), 231

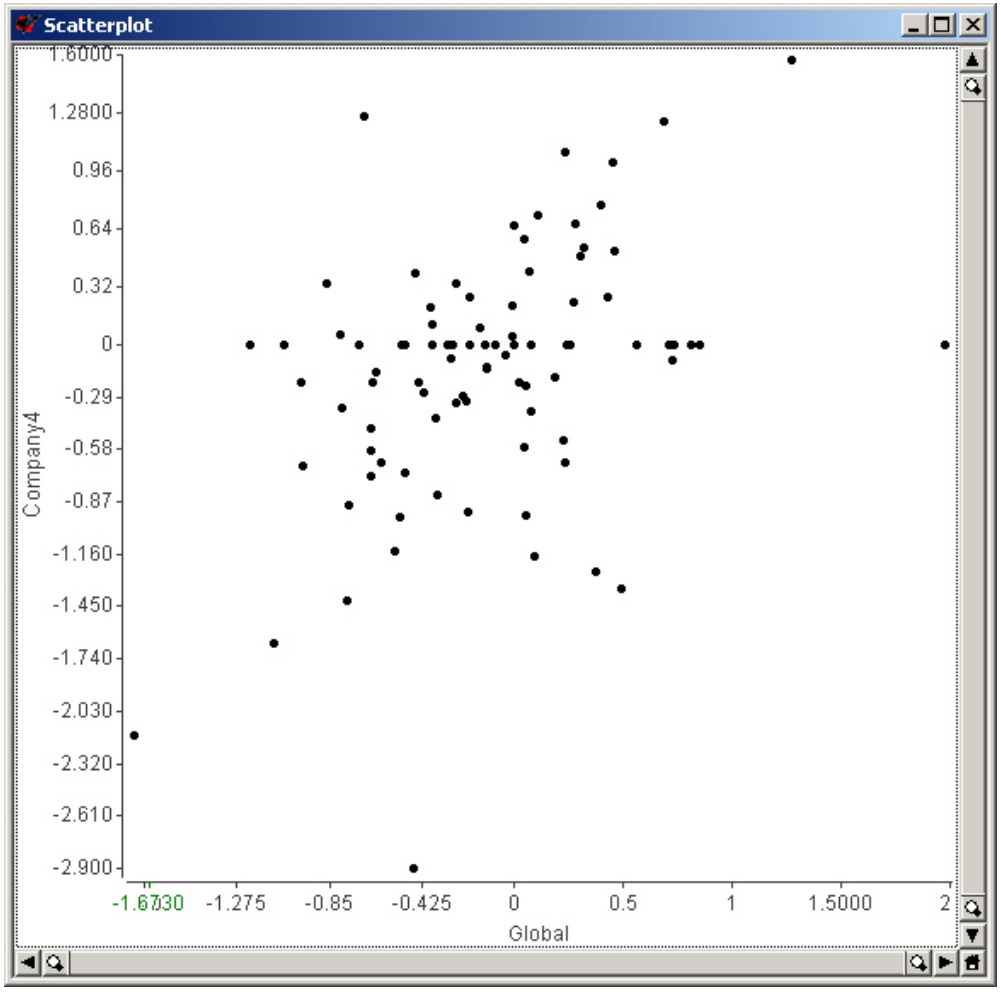
Results: Company 1 (n = 499)



Results: Company 2 (n = 572)



Results: Company 4 (n = 231)



Back to Regression: What about the Rest of the Analysis?

1. Other securely computable statistics
   - $R^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 / \sum_{i=1}^{n}(y_i - \bar{y})^2$
   - $S^2 = (y - X\hat{\beta})^T(y - X\hat{\beta})/(n - p)$
   - Hat matrix $H = X(X^TX)^{-1}X^T$, giving information on outliers
2. Shared synthetic residuals
   - Each agency
     1. Synthesizes predictor values similar to its own
     2. Using *global* regression coefficients, synthesizes residuals associated with its synthetic predictors in a way that mimics the predictor–residual relationship in its own data
   - Agencies share synthetic predictors and residuals via secure data integration

Each agency can assess fit of global model to its data

The Same Idea Works for . . .

Secure Data Integration Data values shareable but sources are not

Secure Contingency Tables Right data structure for large (sparse) table is list of (cell coordinate, cell value) pairs for (only) cells with non-zero values.

- Use secure data integration to build list
- Use secure summation to calculate table entries

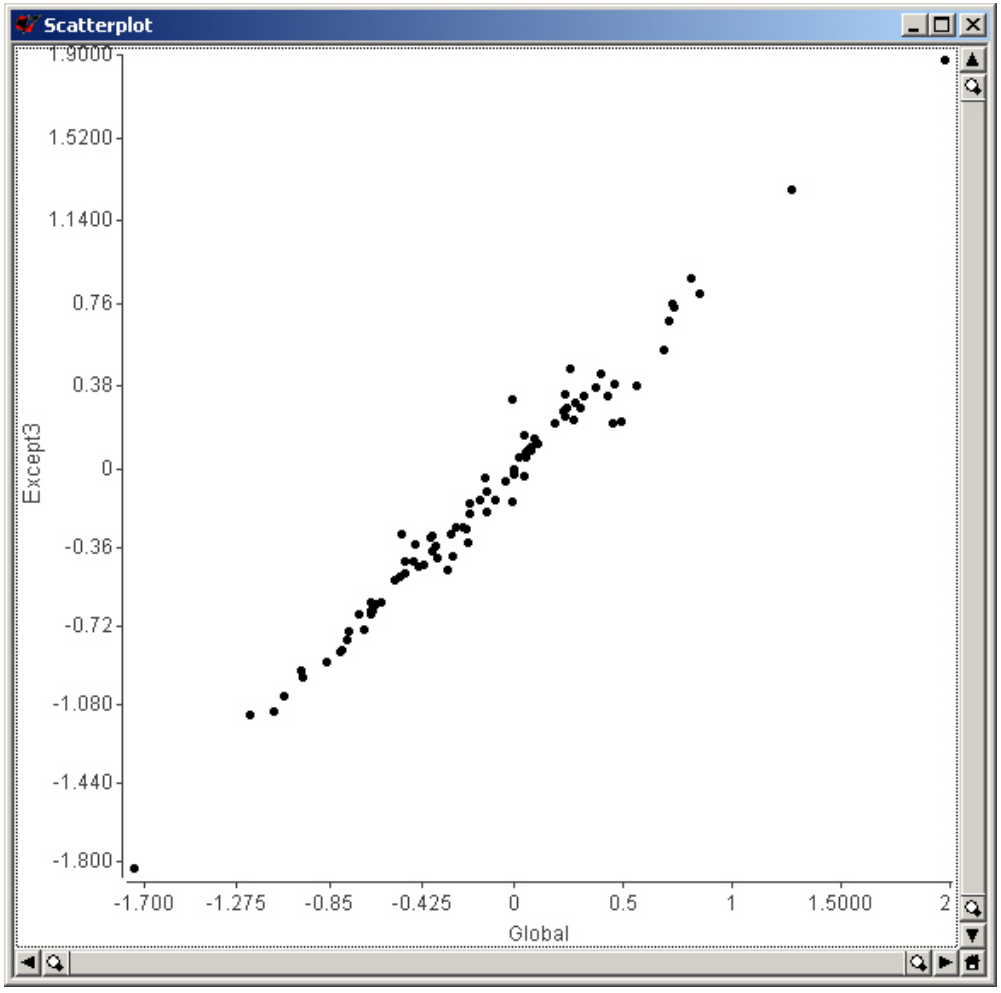
Secure MLE Exponential families with global log-likelihood

$$\log L(\theta, x) = \sum_{\ell=1}^{L} d_\ell(\theta) \left[\sum_{k=1}^{K} \sum_{i \in \text{ Agency } k} c_\ell(x_i) \right]$$

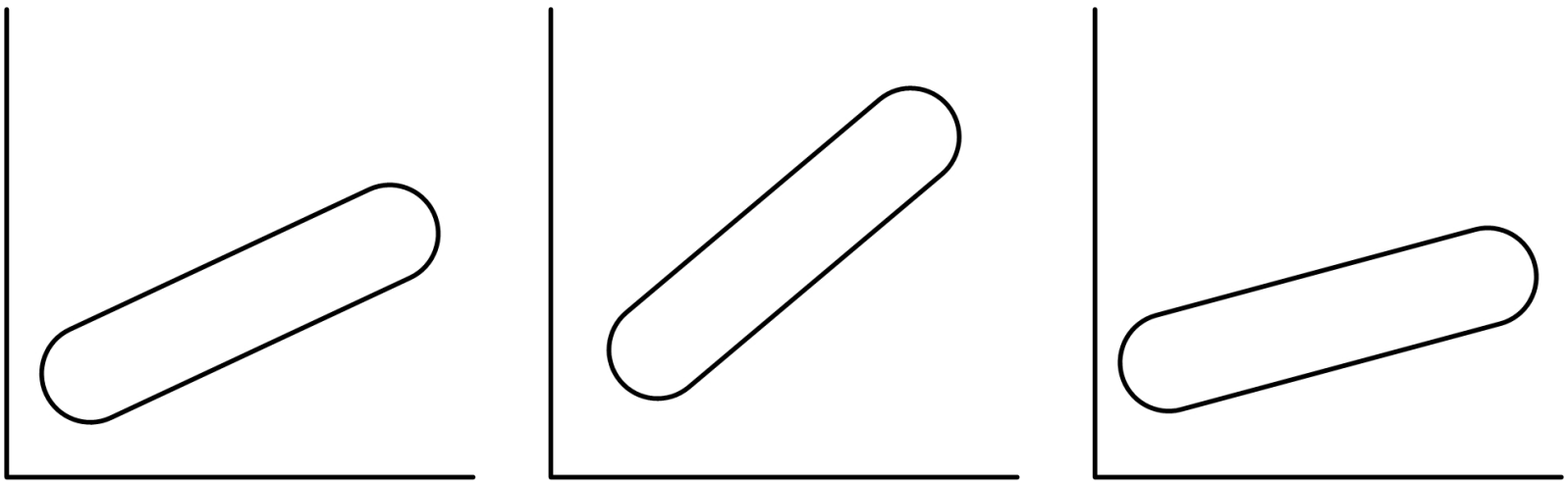
So What's Wrong?

- We can't be good statisticians: analysis must be pre-specified. What about EDA? Visualization?
- Can't protect against dishonesty: if all agencies but one are semi-honest, it can ensure that it gets the right answer and none of the other agencies does or even knows that it doesn't
  - PTTP can reduce incentives, but no ways to detect cheating
- We don't understand risks arising from diversity, e.g.,
  - Unequal database sizes
  - Data heterogeneities (real point: what's the model?)
  - Differential model fit across owners
- No measures—collective or owner-specific—of analysis utility
- Can't (or have only dorky ways to) handle non-additivity. Examples: maximum, sorting, . . .

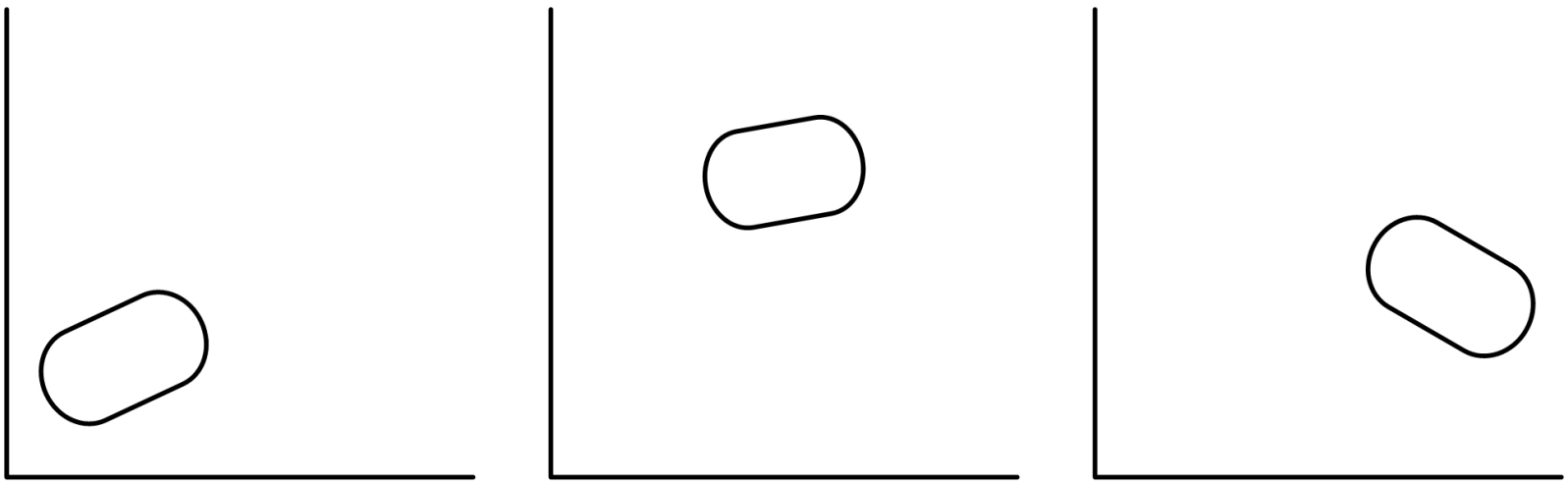
Should Companies 1, 2 and 4 Allow 3 to Participate?



Homogeneous Data



Heterogeneous Data



What Else is Wrong?

- No method to allow agencies to opt out *based on the results of the analysis*
  - Can't do (k, p)-rules in SDL
  - Can do anonymous opt out using (n, p)-rules, but in some cases only post-analysis
- Numerical and algorithmic issues, especially in cases (e.g., iterative computations) where the integrity of the process depends on each machine
- How to perform, and risks associated with, pre-processing: agencies must have attributes in the same units and same order, and must ensure that there are no duplicate records

Set-Up

Agencies A, B, . . . , Ω

Global database $\mathbf{X}$ partitioned vertically among agencies:

$$\mathbf{X} = \left[\mathbf{X}^A \; \mathbf{X}^B \; \cdots \; \mathbf{X}^\Omega\right]$$

Computational need $(p \times p)$-dimensional full data covariance matrix $\mathbf{X}^T\mathbf{X}$ (for regression, . . .)

Goal As little surrender of information as possible

Computing $\mathbf{X}^T\mathbf{X}$

On-diagonal blocks $(\mathbf{X}^A)^T\mathbf{X}^A$ computed by each agency and shared with the others

Off-diagonal blocks $(\mathbf{X}^A)^T\mathbf{X}^B$ computed by pairs of agencies using secure matrix multiplication and shared with the others

Secure Matrix Multiplication Protocol

Step 1 Agency A generates a set of g n-dimensional vectors $\{Z_1, Z_2, \ldots, Z_g\}$ such that $Z_i^T X_j^{\mathrm{A}} = 0$ for all i and j, and sends to agency B the $(n \times g)$-dimensional matrix $\mathbf{Z} = [Z_1\ Z_2\ \cdots\ Z_g]$

Step 2 Agency B computes $\mathbf{W} = \left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^T\right)\mathbf{X}^{\mathrm{B}}$ where $\mathbf{I}$ is an $(n \times n)$-dimensional identity matrix, and sends $\mathbf{W}$ to agency A

Step 3 Agency A calculates $(\mathbf{X}^{\mathrm{A}})^T\mathbf{W} = (\mathbf{X})^T(\mathbf{I} - \mathbf{Z}\mathbf{Z}^T)\mathbf{X}^{\mathrm{B}} = (\mathbf{X}^{\mathrm{A}})^T\mathbf{X}^{\mathrm{B}}$ and shares $(\mathbf{X}^{\mathrm{A}})^T\mathbf{X}^{\mathrm{B}}$ with other agencies

Two extreme cases

- $g = 0$: $\mathbf{W} = \mathbf{X}^{\mathrm{B}}$, so A learns agency B's data exactly
- $g = n - p$: B knows orthogonal complement of $\mathbf{X}^{\mathrm{A}}$ in $\mathbb{R}^n$

Choice of g

Loss of protection to one agency: number of (linearly independent) constraints the other agency has on its data

- Agency A: $\mathrm{LP(A)} = p_A p_B + p_A g$
- Agency B: $\mathrm{LP(B)} = p_A p_B + p_B (n - g)$
- Total, as a function of g:

$$\mathrm{LP}(g) = 2p_A p_B + np_B + (p_A - p_B)g$$

Note: $p_A = p_B$ implies $\mathrm{LP}(g) \equiv 2p_A p_B + np_B$

Inequity: $\mathrm{I}(g) = |\mathrm{LP(A)} - \mathrm{LP(B)}| = |(p_A + p_B)g - np_A|$ is minimized by

$$g^* = \frac{p_A}{p_A + p_B} n$$

Problems

- Analysis must be specified in advance
- Inherent asymmetry: holder of the response loses the most
- Agencies must link records, which
  - Requires common primary key
  - Has severe privacy implications
- Other preprocessing issues: duplicate attributes, incomplete records

Another Problem: Multiple Threats to Record-Level Privacy

- $(\mathbf{I} - \mathbf{Z}\mathbf{Z}^T)$ contains a column with all zeros except for a non-zero constant in one row: *A* learns the value of agency *B*'s data for the data subject in that row
- Attribute that equals zero for all but one data subject
- Sparseness of $\mathbf{X}^A$
- Constraints of the form "gross income $\geq$ net income plus federal tax plus state tax"
- Record in $\mathbf{X}^A$ with dominant attribute values
- High R^2 can imply A has good predictors of B's attributes

Complex Partitions: What We Don't Know

Short answer: Almost everything

Longer but not necessarily more informative answer:

- Computational protocols
- Risks, both old and new (What if knowing who holds what data is risky?)
- Utility measures

References

www.niss.org/dgii/techreports.html