

NISS

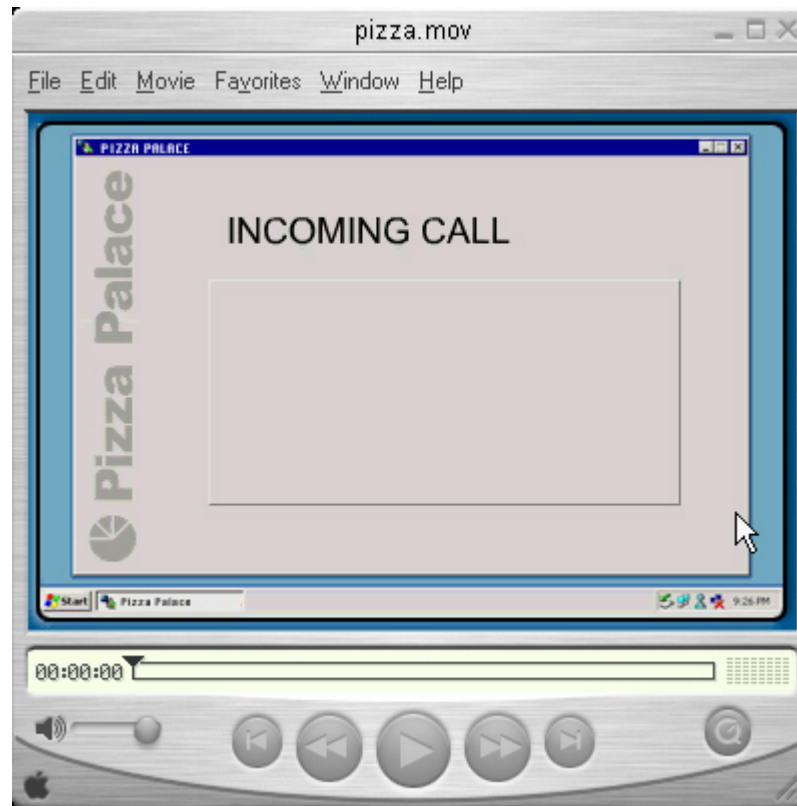
Analysis of Integrated Data without Data Integration

Alan F. Karr

National Institute of Statistical Sciences

karr@niss.org

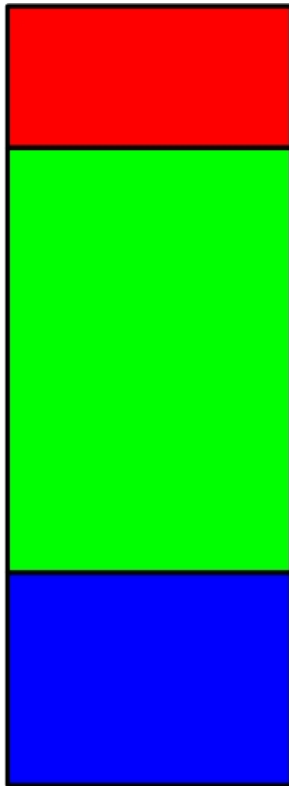
Your Privacy is Threatened!?



Problem Formulation

- Multiple, distributed databases held by different “owners”
 - Government agencies
 - Corporations
- Goals
 - Valid statistical inference on “integrated” database without actually creating it
 - Protect each owner’s data from the other owners
 - [Protect data subjects]
- Constraints
 - No literal sharing of data
 - No trusted third party (human or machine)
 - Semi-honest agencies (more later)

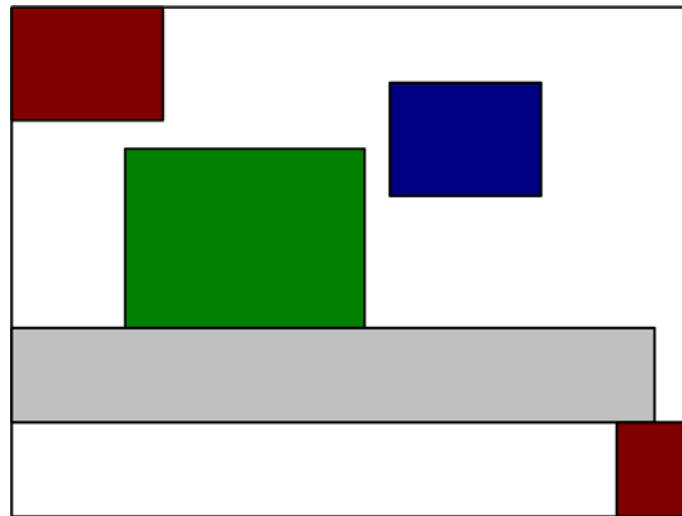
Data Partitioning Models



Horizontal



Vertical



Complicated

The Root:

Secure Multiparty Computation

- Setting
 - Agencies $1, \dots, K$ with values v_1, \dots, v_K
 - Known function f with K arguments
- Goal: Compute $f(v_1, \dots, v_K)$ in such a way that
 - All agency j knows about other agencies' values is what can be deduced from v_j and $f(v_1, \dots, v_K)$
 - Outside parties are not involved
- CS literature
 - Lots of “theorems”
 - Few implementations

The Tool: Secure Summation

- Problem

- Agency k has v_k
- Compute $f(v_1, \dots, v_k) = \sum v_k$

- 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 $R + \sum v_k$, subtract R and share result

Regression for Horizontally Partitioned Data

- Setting: Same data on disjoint sets of subjects
 - y = response
 - X = predictors
- Goal: Perform the regression $y = X\beta + \varepsilon$
including diagnostics
- Constraints
 - No sharing of actual data
 - No trusted third party
 - Semi-honesty

Solution via Secure Summation

- Compute

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

entrywise by secure summation (only $\sim p^2/2$ entries of $X^T X$ need be calculated)

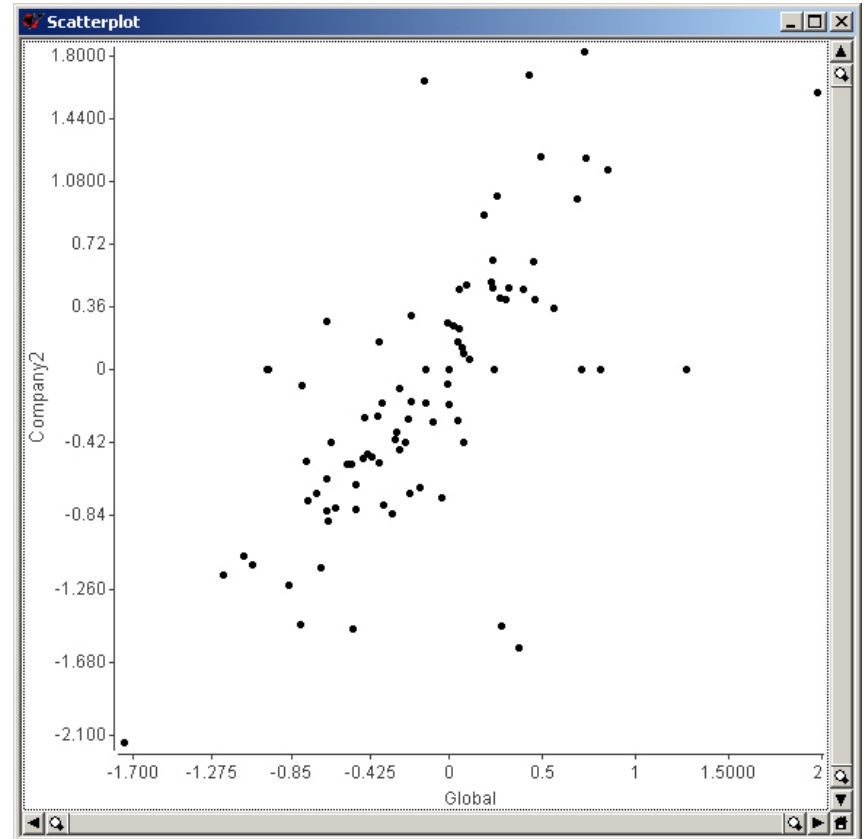
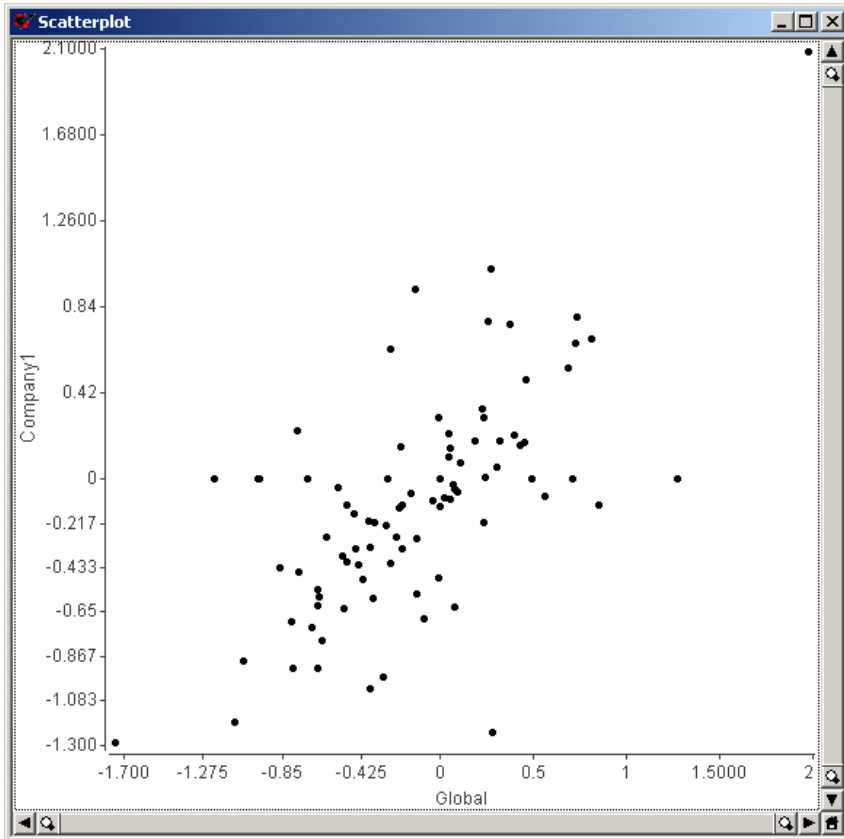
- Share these among agencies; each calculates

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

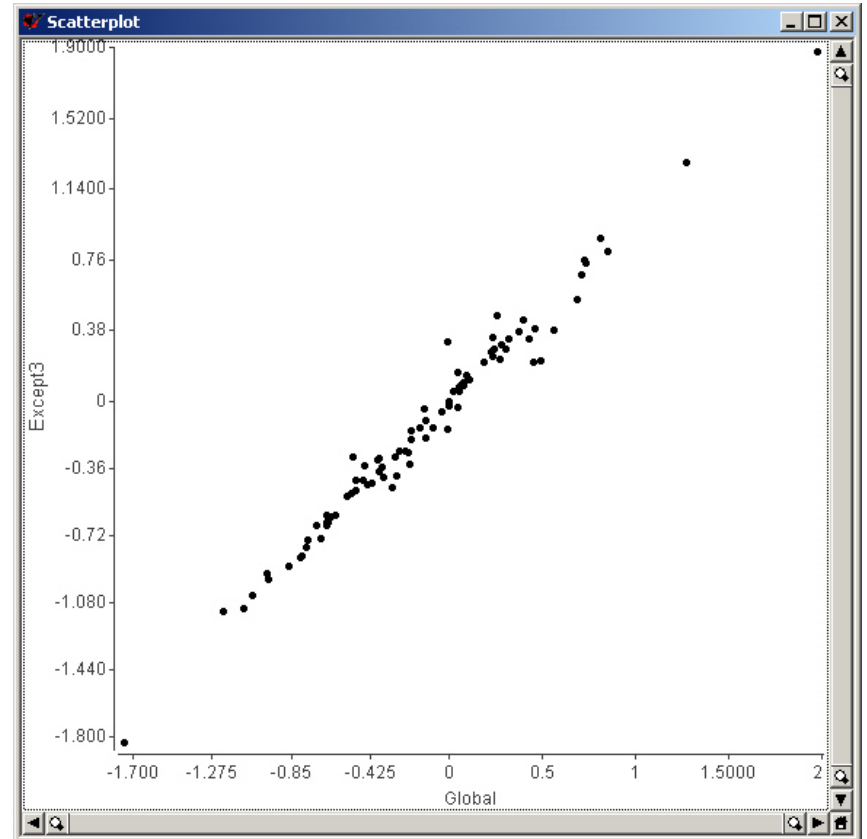
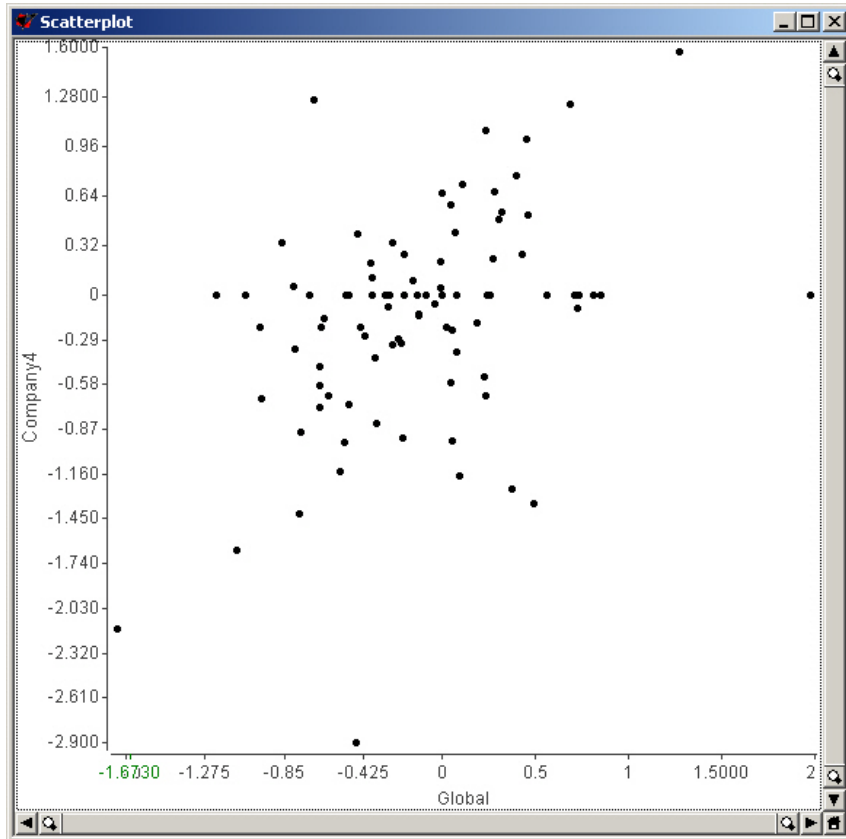
Example: Chemical Data from Multiple Pharmaceutical Manufacturers

- Data
 - 1318 molecules
 - Response: water solubility
 - Predictors
 - 1 constant
 - 90 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



Results—2



Diagnostics

- Securely shared residual statistics
 - R^2
 - S^2
 - Hat matrix $H = X(X^T X)^{-1} X^T$
- Shared synthetic residuals
 - Each agency
 - Synthesizes predictor values *similar to its own*
 - 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*

Secure Contingency Tables

- Key: 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 of coordinates of non-zero cells
 - “Data” are coordinates
- Use secure summation to calculate value for each non-zero cell

Secure MLE

- Assume exponential family:

$$\log f(\theta, x) = \sum_{\ell=1}^L c_{\ell}(x) d_{\ell}(\theta)$$

- Then global log-likelihood is

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

- So, use secure summation on each of L terms

A Closer Look at Semi-Honesty

- Semi-honesty requires agencies to
 - Use correct data
 - Perform agreed-on computations properly

But, allows them to retain results of intermediate computations

- Is there an advantage to not being semi-honest?
- In original protocol for regression for HP data, agency j knows
 - Its own $(X^j)^T X^j$ and $(X^j)^T y$
 - Global $X^T X$ and $X^T y$

Problem Scenarios

- Agency j puts in 0 instead of $(X^j)^T X^j$ and $(X^j)^T y$:
 - Calculated global regression = complementary regression for agencies other than j
 - Agency j can add $(X^j)^T X^j$ and $(X^j)^T y$ to get the correct global regression
 - Other agencies have correct answer to wrong question
- Agency j puts in junk instead of $(X^j)^T X^j$ and $(X^j)^T y$:
 - Calculated global regression = garbage
 - Agency j can subtract junk, add $(X^j)^T X^j$ and $(X^j)^T y$ and end up with the correct global regression
 - Other agencies have garbage and don't know it

Partially Trusted Third Party

- Agencies give up some knowledge to the PTTP to protect themselves from one another
- Operationally, the PTTP is a data-less agency that performs and shares the result of a particular calculation
- Works when result is $f(S_1, S_2)$, where S_1 and S_2 are statistics calculated using secure [summation or ...]

PTTP for Secure Regression

- Agency $K+1$ that
 - Has no data
 - Initializes calculation of $X^T X$ and $X^T y$ with random numbers
 - After agencies $1, \dots, K$ contribute,
 - Calculates $X^T X$ and $X^T y$ by removing random numbers
 - Calculates $\hat{\beta} = (X^T X)^{-1} X^T y$
 - Shares $\hat{\beta}$ with the other agencies

Does it Work?

- Advantages
 - Removes one incentive to cheat
 - Especially compatible with star topology to handle communication
 - Prevents collusion, because agencies unaware of order
 - Detects if agency puts in 0
- Disadvantages
 - PTPP knows more than the agencies

Current Questions

- How much less does PTTP reveal?
 - Under original protocol, agency j knows complementary regression exactly
 - Under PTTP, what does agency j know about $\hat{\beta}_{-j}$?
- Is PTTP stable?
 - If agency j puts in false data, does it make the other agencies worse off than it makes itself?
- Can PTTP handle
 - R^2 and standard errors: yes
 - Diagnostics: some
 - Other analyses: ???
- Is PTTP saleable?

References

(Available at www.niss.org/dgii/techreports.html)

- A.F. Karr, X. Lin, J. P. Reiter and A. P. Sanil (2005). Secure Regression on Distributed Databases. *J. Computational and Graphical Statist.* 14(2) 263-279.
- A. F. Karr, X. Lin, J. P. Reiter and A. P. Sanil (2004). Analysis of Integrated Data without Data Integration. *Chance* **17(3)** 26-29.
- A. F. Karr, X. Lin, J. P. Reiter and A. P. Sanil (2004). Privacy Preserving Analysis of Vertically Partitioned Data using Secure Matrix Products. Submitted to *J. Official Statist.*
- A. F. Karr, X. Lin, J. P. Reiter, A. P. Sanil (2005). Secure statistical analysis of distributed databases. In *Statistical Methods in Counterterrorism*, D. Olwell and A. G. Wilson, eds. (to appear).
- J. P. Reiter, A. F. Karr, C. N. Kohnen, X. Lin, and A. P. Sanil (2004). Secure Regression for Vertically Partitioned, Partially Overlapping Data. *ASA Proceedings*.
- A. P. Sanil, A. F. Karr, J. P. Reiter and X. Lin (2004). Privacy Preserving Regression Modeling via Distributed Computation. *Proc. Tenth ACM SIGKDD* 677-682.