

# NISS

## Regression on Distributed Databases via Secure Multi-Party Computation

Alan F. Karr  
National Institute of Statistical Sciences  
karr@niss.org

March 21, 2005

# Outline

- Problem formulation
- Secure multi-party computation
- Horizontally partitioned data
- Vertically partitioned data
- What we don't know

# Formulation

- Related databases held by multiple “agencies”
- Actual data integration impossible
  - Confidentiality
  - Regulation
  - Proprietary data
  - Scale of the data
- Agencies wish to perform sound statistical analyses on integrated data—regression, classification, data mining, ...

# Constraints

- No trusted third party (human or machine)
- Cooperating, semi-honest agencies
  - Use true data
  - Follow agreed-on protocols
  - Can retain results of intermediate computations
- No collusion

# Prototype Problem: Regression

- Horizontally partitioned data
  - Same attributes
  - Different subjects
  - Example: state databases on K-12 students
- Vertically partitioned data
  - Different attributes
  - Same subjects
  - Example: BLS (employers), NCES (schools), NCHS (health)

# Secure Multi-Party Computation

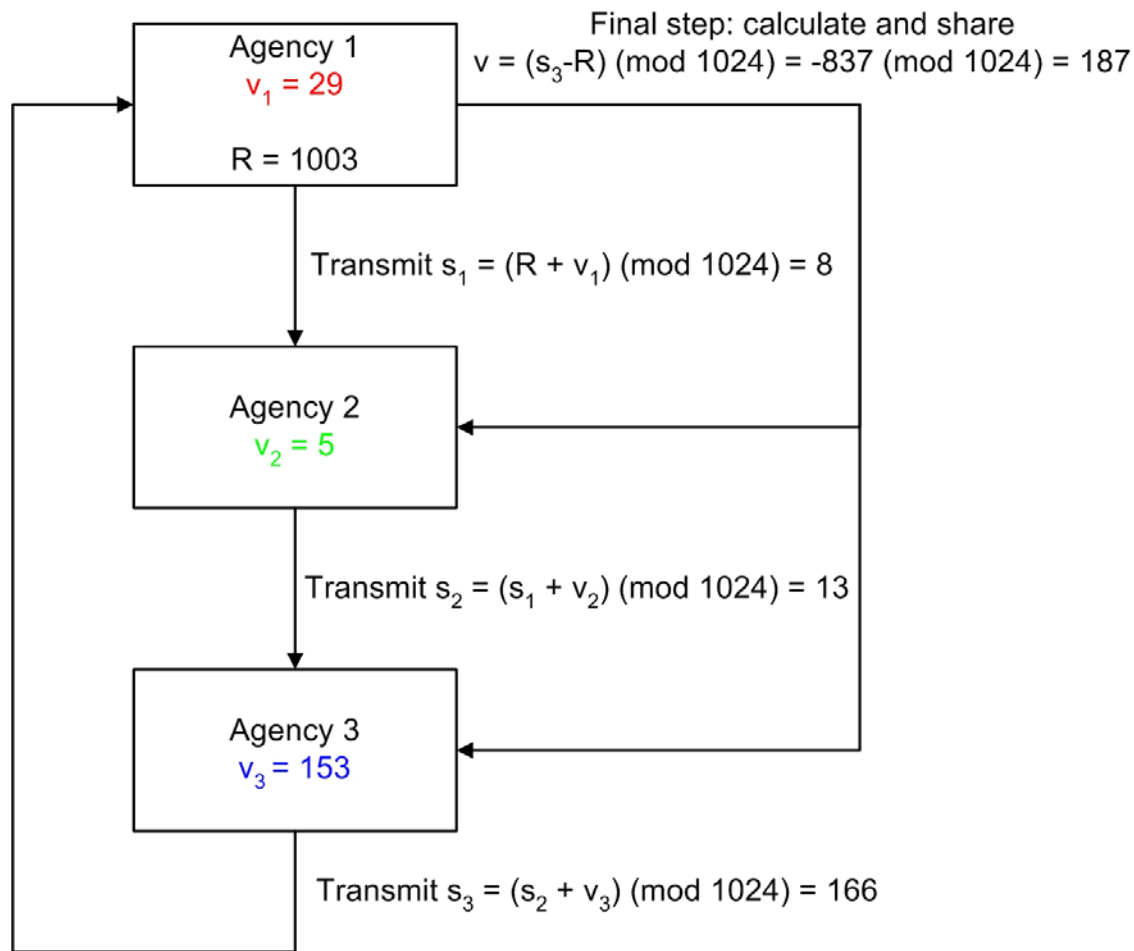
- Agency  $i$  has  $A_i$ ,  $i=1, \dots, K$ , and they want to compute  $f(A_1, \dots, A_K)$  for known  $f$  in such a way that agency  $i$  learns no more about  $\{A_j : j \neq i\}$  than can be deduced from  $A_i$  and  $f(A_1, \dots, A_K)$
- Lots of CS generalities
  - Secure XOR
  - “Any computation can be reduced to XOR”
- Not many functioning algorithms, let alone systems

# Secure Summation:

$$f(A_1, \dots, A_K) = A_1 + \dots + A_K$$

- Agency 1
  - Generate enormous random number  $R$
  - Transmit  $R + A_1$  to party 2
- Agency 2
  - Add  $A_2$
  - Transmit  $R + A_1 + A_2$  to party 3
- ...
- Agency 1
  - Receive  $R + \sum A_k$
  - Subtract  $R$
  - Share the result

# Secure Summation for $K = 3$



# Approaches to Secure Regression

- *Secure data integration*
  - Create pooled database without revealing the sources of the records
- *Secure analyses*
  - Locally compute and securely share “sufficient statistics”

# Secure DI: Version 1

- Round 1
  - Agency 1 puts in only synthetic data
  - Agency 2, ..., K puts in at least 5% of its real data; optionally, puts in synthetic data; randomly permutes order of records
- Rounds 2, ..., 20
  - Agency 1, ..., K puts in at least 5% of its real data; optionally, puts in synthetic data; randomly permutes order of records
- Round 21
  - Agency 1 puts in any remaining real data; removes its synthetic data
  - Each agency 2, ..., K removes its synthetic data
- Problems
  - Retained intermediate computations: agency 3 can identify the real data it receives from agency 2 in Round 1
  - Vulnerable to *bad or good* synthetic data
  - Doesn't deal with attribute values that are informative about source

# Secure DI: Version 2

- Stage 1 agency  $a_1$ 
  - Initializes database with some synthetic data and at least one real data record
  - Picks stage 2 agency  $a_2$  randomly and sends database and indicator vector  $d$ , where  $d_i = 1$  ( $i$  has data left)
- While one or more agencies have data left, stage  $j$  agency  $a_j$ 
  - Adds at least one real record and optional synthetic data
  - Sets  $d_{a_j} = 0$  if it has no data left
  - Chooses  $a_{j+1}$  randomly from agencies with data left
- Final stage
  - Agencies remove synthetic data

# Secure Regression for Horizontally Partitioned Data

- Setting
  - Horizontally partitioned data: agencies have same data on disjoint sets of subjects
  - $Y$  = response,  $X$  = predictors
- Goal: Perform the regression

$$Y = X\beta + \varepsilon$$

*including diagnostics*

- Use secure combination of locally computed sufficient statistics

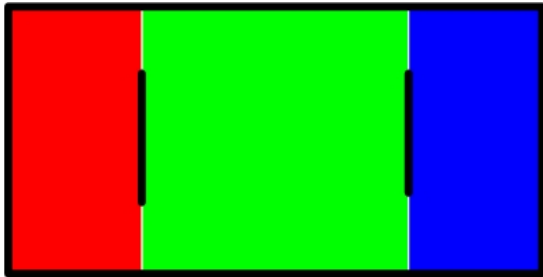
# Secure Regression without DI

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

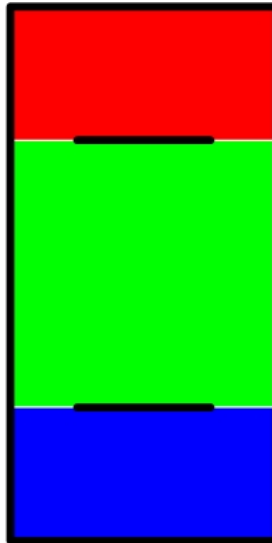
$$X^T Y = \sum_{j=1}^K (X^j)^T Y^j$$

- Compute each entrywise by *secure summation*
  - Only  $\sim p^2/2$  entries of  $X^T X$  need be calculated
- Share among agencies
- Each agency calculates  $\hat{\beta} = (X^T X)^{-1} X^T Y$

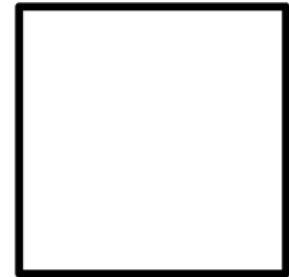
# Pictorial View



$$X^T: p * (n_1 + n_2 + n_3)$$



$$X: (n_1 + n_2 + n_3) * p$$



$$X^T X: p * p$$

# Example: Boston Housing Data

- Data: 506 cases, representing towns, partitioned into, Agency 1: 172, Agency 2: 182, Agency 3: 152
- Model: Response = housing value, 3 predictors + constant
  - CRIME
  - IND[USTRIALIZATION]
  - DIST[ANCE] from employment centers

Regression	$\hat{\beta}_{\text{CONST}}$	$\hat{\beta}_{\text{CRIME}}$	$\hat{\beta}_{\text{IND}}$	$\hat{\beta}_{\text{DIST}}$
Global	35.505	-0.273	-0.730	-1.016
Agency 1	39.362	-8.792	-0.720	-1.462
Agency 2	35.611	2.587	-0.896	-0.849
Agency 3	34.028	-0.241	-0.708	-0.893

# Diagnostics

- Securely shared residual statistics
  - $R^2$ ,  $S^2$ , ...
- 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*
  - Shares synthetic predictors and residuals via secure DI
  - Can assess
    - Fit of global model to its own data
    - Global fit of global model

# Secure Regression for Vertically Partitioned Data

- Setting
  - Vertically partitioned data: agencies have disjoint sets of attributes for the same subjects
  - $Y$  = response,  $X$  = predictors
- Goal: Perform the regression

$$Y = X\beta + \varepsilon$$

*including diagnostics*

- Use secure combination of locally computed sufficient statistics

# Approaches

- Case 1: only one agency holds the response
  - Use secure matrix products to compute full data covariance matrix
- Case 2: all agencies hold the response
  - Use Powell's algorithm to solve

$$\hat{\beta} = \arg \min_{\beta} \|X\beta - Y\|^2$$

# What We Don't Know

- How to build functioning servers for secure DI and secure regression
- With few exceptions, how to do other kinds of analyses
  - Example: classification
  - Exception: contingency tables
- The SDL implications
- How to handle uncertain record linkage

# Some References

([www.niss.org/dgii/technicalreports.html](http://www.niss.org/dgii/technicalreports.html))

- A. F. Karr, X. Lin, J. P. Reiter, A. P. Sanil (2004). Analysis of integrated data without data integration. *Chance* **17(3)** 26-29
- A. F. Karr, X. Lin, J. P. Reiter, A. P. Sanil (2004). Secure regression on distributed databases. *J Computational and Graphical Statist.* (to appear)
- A. F. Karr, X. Lin, J. P. Reiter, 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 regression on distributed databases. To appear in *Statistical Methods in Counterterrorism*, D. Olwell and A. G. Wilson, eds. ASA/SIAM Series on Statistics and Applied Probability.
- A. P. Sanil, A. F. Karr, X. Lin, J. P. Reiter (2004). Privacy preserving regression modelling via distributed computation. *Proc. Tenth ACM SIGKDD Internat. Conf. on Knowledge Discovery and Data Mining* 677-682.