

NISS

Regression on Distributed Databases via Secure Multi-Party Computation

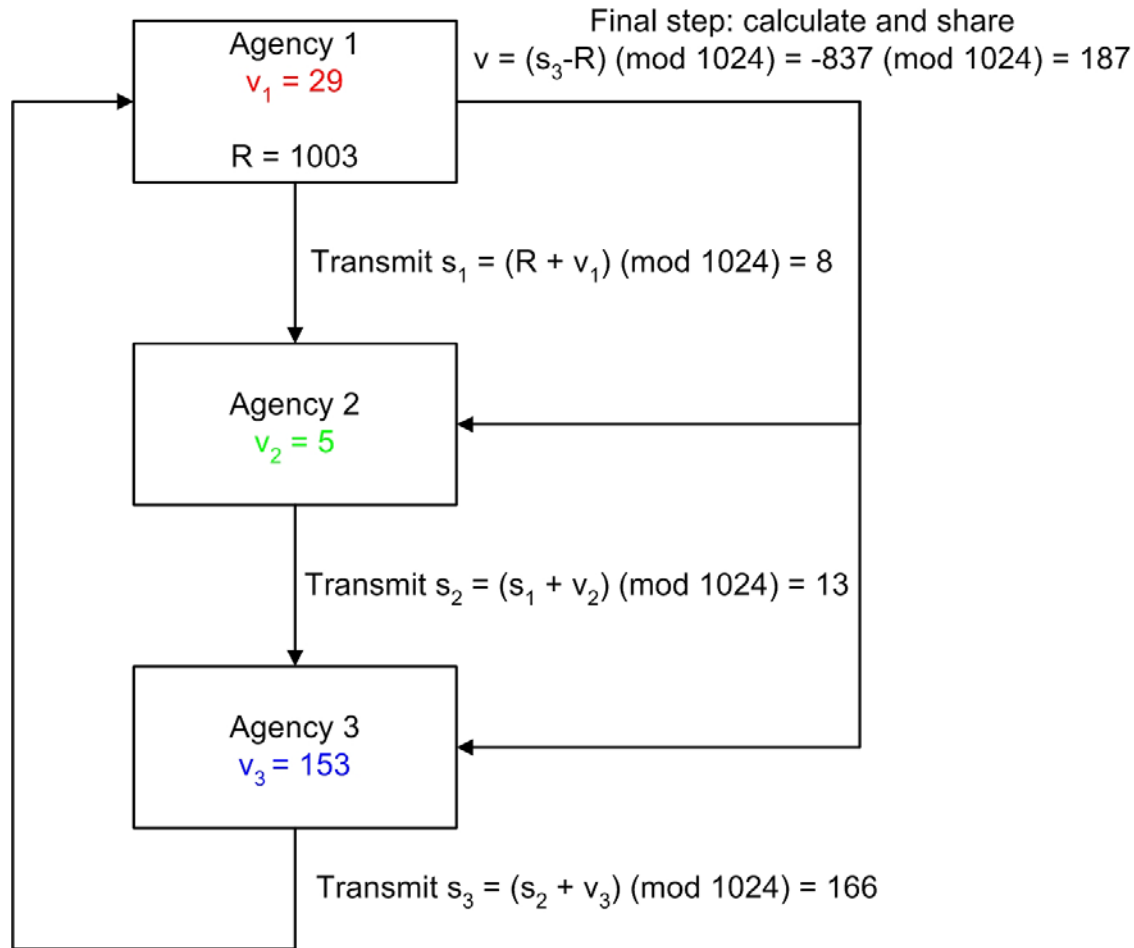
Alan F. Karr, Xiaodong Lin, Ashish P. Sanil
National Institute of Statistical Sciences

Jerome P. Reiter
Duke University

Problem Summary

- Goal: Statistical analyses that “integrate” data stored in multiple, distributed databases
- Impediment: barriers to actually integrating the databases
 - Confidentiality
 - Regulation
 - Scale
- Solution: Use secure multi-party computation

Example: Secure Summation



Secure Linear Regression

- $K > 2$ agencies with horizontally partitioned data
 - Same attributes for different sets of data subjects
 - Agency j has n_j subjects
- Usual linear model $y = X\beta + \varepsilon$
 - p predictors x_1, \dots, x_p ($x_1 = 1$)
 - Response y
- Least squares estimators

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

The Math

1

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

- Compute entrywise by secure summation (only $\sim p^2/2$ entries need be calculated)
- Share among agencies

2

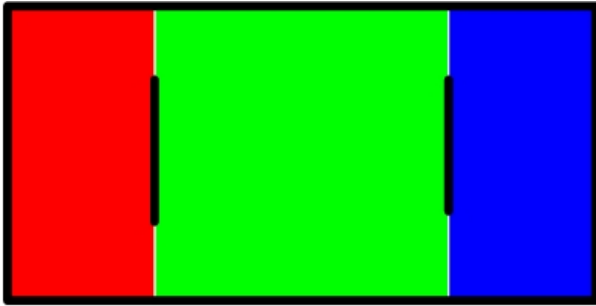
$$X^T y = \sum_{j=1}^K (X^j)^T y^j$$

- Compute entrywise by secure summation
- Share among agencies

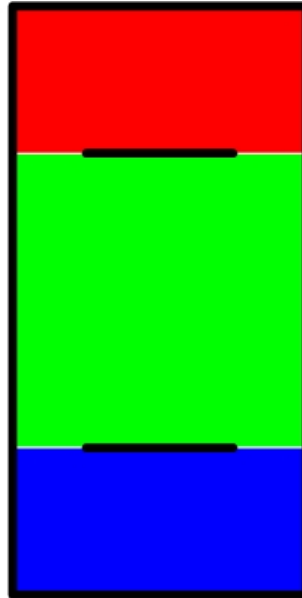
3

All agencies calculate $\hat{\beta}$

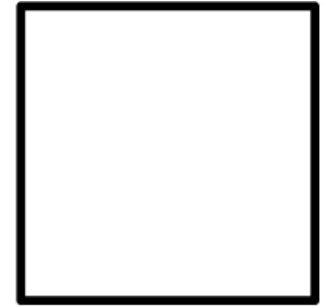
The Picture



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



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



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

Example: Boston Housing Data

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

Boston Housing Data Results

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

Other Issues

- Diagnostics
 - Computable from local statistics using secure summation.
 - Example: R^2
 - Securely pooled synthetic residuals
- Opting out
 - Can be done securely
- Variations
 - Vertically partitioned data
 - More complex partitioning
 - Secure record linkage?????
 - Categorical data: secure contingency tables