NISS

Secure Analyses of Distributed Data

Alan Karr
March 4, 2004

Context

- Related databases held by multiple parties ("agencies")
  - Government agencies
  - Corporations (e.g., pharmaceutical companies)
- Actual data integration impossible
  - Law
  - Proprietary data
  - Data size
- Wish to perform statistical analyses on integrated data
  - Data mining
  - Regression
  - …

Constraints

- No trusted third party (human or machine)
- Cooperating agencies
  - Want to perform the analyses
- Semi-honest agencies
  - Use true data
  - Follow agreed on protocols
  - Can retain results of intermediate computations

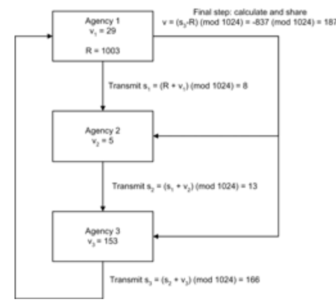
Data Partitioning

- Horizontal
  - Agencies have same data on disjoint sets of subjects
  - Example: state-level education data
- Vertical
  - Agencies have disjoint sets of attributes on the same subjects, and "clean" record linkage is possible
  - Example: IRS, NCES, NCHS
- Mixed

Secure Summation

- Problem
  - Party k has a_k
  - Compute Σa_k without revealing any of the a_k to others, and without trusted third party (human or machine)
- Solution
  - Party 1: generate enormous random number R, and transmit $R + a_1$ to party 2
  - Party 2: Add a_2, transmit $R + a_1 + a_2$ to party 3
  - ...
  - Party 1 receive $R + \Sigma a_k$, subtract R and share result

Secure Summation: Pictorial View



Confidentiality-Preserving Association Rules

- Problem setting: multiple, identical databases with different owners
- Goal: find item pairs *(i,j)* with *global* (across all the databases) association rule support exceeding threshold *s*
- Constraint: protect
  - Data items
  - Database sizes N_k
  - Support $S_k = C_k(i,j)/N_k$ at each site
- Answer: Secure summation with $a_k = C_k(i,j) - sN_k$ to compute

$$1\left(\sum_k C_k(i,j) - s\sum_k N_k \geq 0\right)$$

Secure Regression

- Setting: horizontally partitioned data
  - Y = response
  - X = predictors
- Goal: Perform ordinary linear regression, *including diagnostics*

Approaches

- Secure data integration
  - Create integrated database in which no agency can recognize the source of data other than its own
- Secure combination of local computations
  - Compute $(X^TX)^{-1}X^TY$ using secure summation
  - Diagnostics via
    - Securely shared local computations
    - Securely integrated synthetic residuals

Secure DI: Version 1

- Round 1
  - Agency *1*
    - Puts in only synthetic data
  - Each 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
  - Each 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
  - In round 1, agency 3 receives
    - Synthetic data from agency 1
    - Real and synthetic data from agency 2
  - By comparing with final database, agency 3 can identify the real data from agency 2
- Vulnerable to poor synthetic data
- Vulnerable to good synthetic data

Secure DI: Version 2

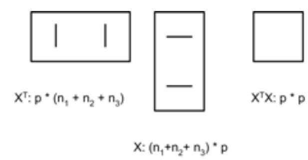
- Stage 1 agency a_1
  - Initialize database with
    - Some synthetic data
    - At least real data record
  - Pick stage 2 agency a_2 randomly and send database and indicator vector d ($d_i = 1$(*i* has data left))
- While two or more agencies have data left,
  - Stage *j* agency a_j
    - Adds at 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

Regression via Secure DI

- Use Version 2 to create and share integrated database
- Each agency can run whatever analyses it wants

Secure Regression without DI

- Model: $Y = X\beta + \varepsilon$
- Least squares estimates
  - Compute X^TX and X^TY entrywise via secure summation
  - All agencies can then compute $\hat{\beta} = (X^T X)^{-1} X^T Y$



Diagnostics via Securely Shared Residual Statistics

- $R^2 = \dfrac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$
- $S^2 = \dfrac{(y - X\hat{\beta})^T (y - X\hat{\beta})}{n - p}$
- Outliers via $H = X(X^T X)^{-1} X^T$

Diagnostics via Shared Synthetic Residuals

- Each agency
  - Simulates predictor values
  - Using shared regression coefficients, simulates residuals associated with its synthetic predictors
- Agencies share synthetic residuals via secure DI

Problems

- Other forms of data
  - Secure method for integrated contingency tables
  - Text, images, …
- Other analyses
- Risk-utility characterization
  - Disclosure risk = ???
  - Data utility = ???
- Compare what is revealed to what has to be revealed

Vertically Partitioned Case

- All agencies have data on same subjects
  - Common primary key
- Agencies “own” disjoint sets of attributes
  - If there are attributes in common, they agree
- Complete data

What We Can Do

- Compute least squares estimators
  - Approach 1: Use secure matrix product to compute “off-diagonal” blocks in covariance matrix
  - Approach 1: Use Powell’s method to solve the quadratic optimization problem

What We Can’t Do

- Diagnostics
- Characterize asymmetries
  - Response holder
  - Calculation of covariance matrix
- Derive any optimality properties

And There's More …

- Incomplete data case
  - Agencies (uniquely) "own" attributes
  - For each subject, each agency has either all or none of its attributes for that subject