NISS

# Secure Data Integration

Alan F. Karr
karr@niss.org

# Outline

- Three basic operations
  - Secure summation
    - Application to secure association rules
  - Secure Boolean
    - Application to secure opt out
  - Secure data pooling
    - Application to categorical data
- Application
  - Secure (horizontally partitioned) regression

# Basics

- Setting
  - Parties with multiple, identical (!?) databases
- Goal
  - Secure computation, analysis, data pooling, … without revealing to any party anything other than what can be deduced by “subtraction” from final result
- Assumptions
  - More than 2 parties
  - Cooperating parties

# Secure Summation

- Problem
  - Party $k$ has $a_k$
  - Compute $\Sigma a_k$ 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

# Example: Association Rules

- D = database, a “flat file”
  - Rows = cases
  - Columns = categorical attributes
- A, B subsets of D, typically defined by “attribute = value”
- Association rule A ⇨ B: to what extent does A imply B?
- Quantifications
  - Confidence: $C = P(B|A) = \#(A \cap B) / \#(A)$
  - Support: $S = P(B \cap A) = \#(A \cap B) / \#(D)$

# Association Rules: Example

- CPS8D data
  - 48,842-element excerpt from 1993 CPS
  - 8 categorical attributes: Age, EmplType, Educ, MarStat, Race, Sex, Hours, Income
  - 2880 (1695 non-zero) cells in full contingency table

Salary x Age

| | <25 | 25-55 | >55 |
|---|---|---|---|
| <$50K | 8339 | 23912 | 4904 |
| >=$50K | 93 | 9629 | 1965 |

Confidence given Age

| | <25 | 25-55 | >55 |
|---|---|---|---|
| <$50K | 0.99 | 71 | .71 |
| >=$50K | .01 | .29 | .29 |

# Secure Association Rules

- Problem
  - Find item pairs *(i,j)* with *global* (across all the databases) association rule support exceeding threshold *s*
  - Protect
    - Data items (no shared data)
    - Database sizes $N_k$
    - Support $C_k(i,j)/N_k$ at each site
- Solution: use secure summation to compute

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

# Secure Boolean

- Problem
  - Determine whether “yes/no” vote is unanimous without revealing
    - No voters
    - Number of no votes
- Solution ($K$ = number of parties)
  - Party 1 creates and sends to Party 2 $p_1$ = product of several (large) primes
  - Party 2 forms and sends on
    - $p_2 = p_1$ * 1 prime if “yes”
    - $p_2 = p_1$ * several primes if “no”
  - …
  - Party 1 ends up with $p_K$
    - Divide by $p_1$ and factor result into primes (needs good algorithm)
      - If $K - 1$ factors, all other votes were “yes”
      - Otherwise, there was a “no” vote
    - Informs others of final result

# Example: Secure Opt Out

- Problem
  - For pooled data, allow anyone to prevent pooling whose database size $N_k$ exceeds $p$% of the total
- Solution
  - Secure summation to compute $N = \sum N_k$
  - Secure Boolean with yes = $N_k \leq pN$

# (Pretty) Secure Data Pooling

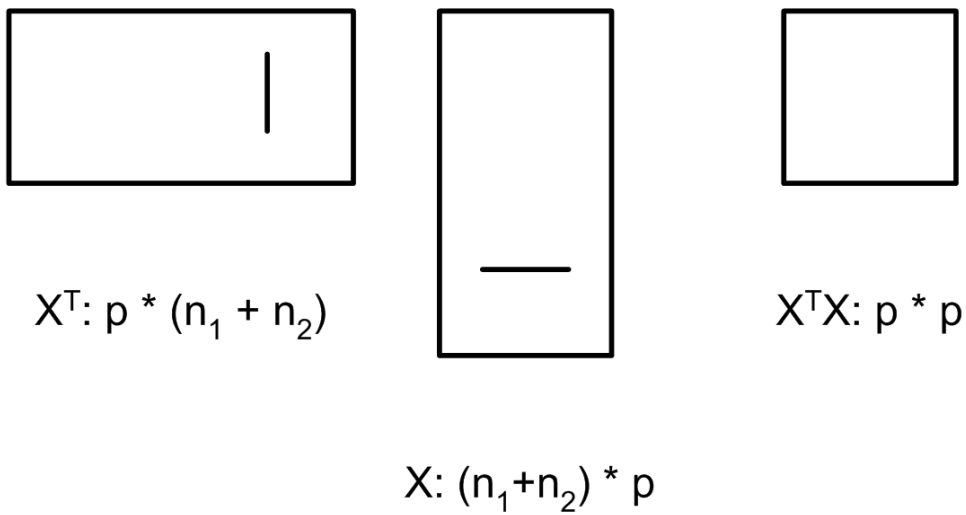
- Problem
  - Build pooled database without revealing to any party the source of “other” elements
- Solution
  - Use secure summation to determine $N = \Sigma N_k$,
  - Party 1 create and send to Party 2 list containing
    - *(100 *) N* “fake” data records (generated via SDL?)
    - At least 5% of his/her real records
  - Party 2 add at least 5% of her/his records to list, and send to Party 3
  - … (at most 20 iterations)
  - Party 1 remove the “fake” records and share pooled data set

# Secure Pooling for Categorical Data

- Solution: build contingency table
  - Unscalable but completely secure version
    - Use secure summation to compute the associated contingency table in array form, cell-by-cell
  - Scalable but less secure version
    - Build sparse representation (coordinates, count)
    - Use secure data pooling to create list of non-zero cells
    - Use secure summation to compute values of non-zero cells

# Secure Regression

- Parameter estimation
  - Compute $X^TX$ and $X^TY$ entrywise via secure summation; share results; everyone can compute $\hat{\beta} = (X^T X)^{-1} X^T Y$




- Diagnostics: under construction