Secure Federated Feature Selection

PRESENTER:
Lun Wang, Qi Pang

BACKGROUND: we propose the first secure federated χ2 -test protocol FED-χ2 . To minimize both the privacy leakage and the communication cost, we recast χ2 -test to the second moment estimation problem and thus can take advantage of stable projection to encode the local information in a short vector. As such encodings can be aggregated with only summation, secure aggregation can be naturally applied to hide the individual updates. We formally prove the security guarantee of FED-χ2 that the joint distribution is hidden in a subspace with exponential possible distributions. Our evaluation results show that FED-χ2 achieves negligible accuracy drops with small client-side computation overhead. In several real-world case studies, the performance of FED-χ2 is comparable to the centralized χ2 -test.

BUILDING BLOCK: STABLE PROECTION

Function ENCODE(P, u_i):
  return $\mathbf{P} \times \mathbf{u}_i$
Function GEOMETRICMEANESTIMATOR(e):
  ▷ ℓ: Encoding size.
  $\hat{d}_{(2),gm} \leftarrow \frac{\prod_{k=1}^{\ell} |\mathbf{e}_k|^{2/\ell}}{(\frac{2}{\pi}\Gamma(\frac{2}{\ell})\Gamma(1-\frac{1}{\ell})\sin(\frac{\pi}{\ell}))^{\ell}}$
  return $\hat{d}_{(2),gm}$
Function DECODE(e):
  return GEOMETRICMEANESTIMATOR(e)

METHODOLOGY

STEP 1: Recast χ2 –test as a second frequency moments estimation problem.

STEP 2: Encode the second frequency moments information using stable projection.

STEP 3: Use secure aggregation to sum the clients' updates.

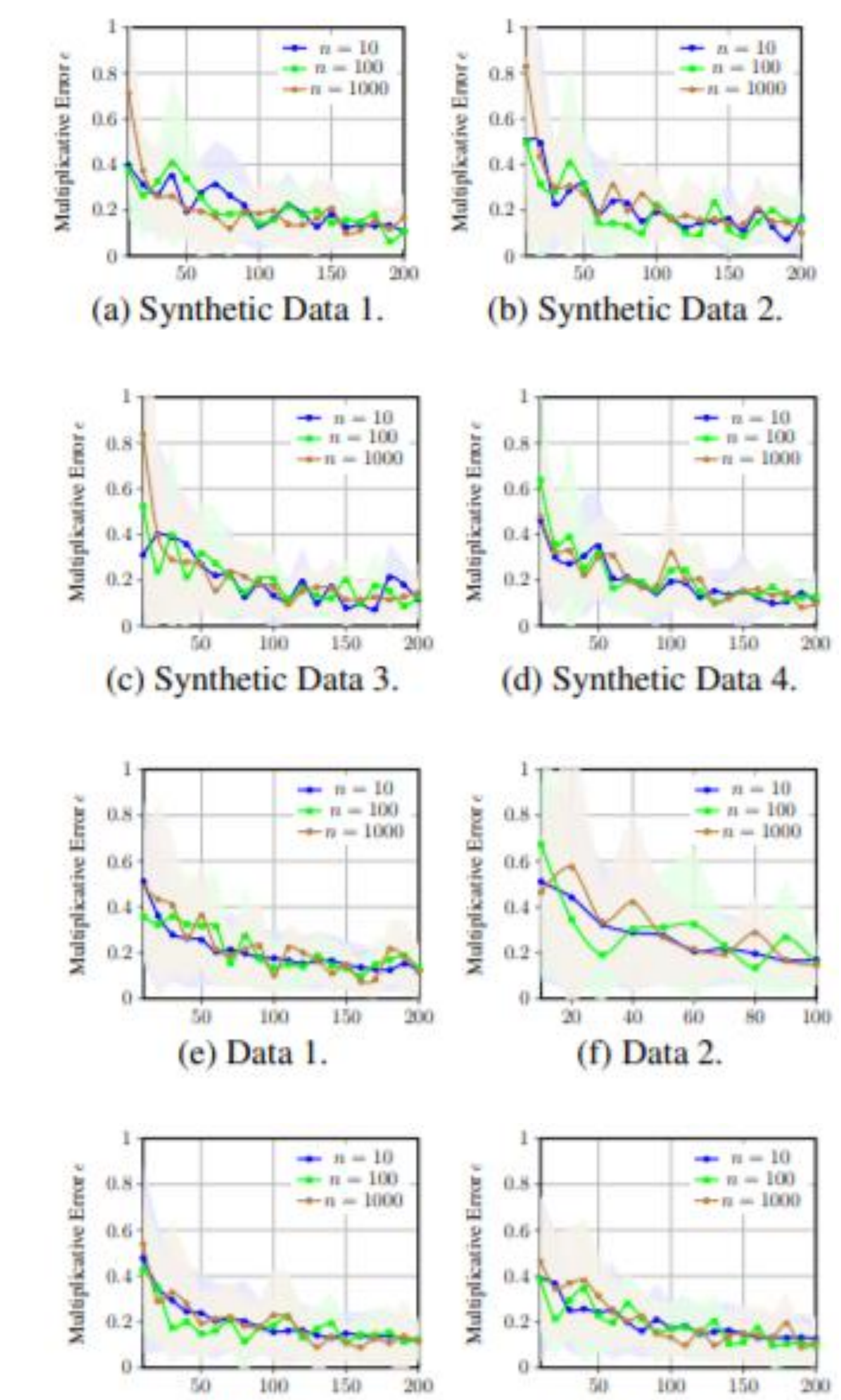
we propose the first secure federated χ2-test protocol FED-χ2.

Algorithm 2: FED-χ^2: secure federated χ^2-test. SECUREAGG is a remote procedure that receives inputs from the clients and returns the summation to the server. INITSECUREAGG is the corresponding setup protocol deciding the communication graph and other hyper-parameters.

Round 1: Reveal the marginal distribution
  INITSECUREAGG(n)   // n is the client number.
  for $x \in [m_x]$ **do** v_x = SECUREAGG($\{v_x^{(i)}\}_{i\in[n]}$)
  for $y \in [m_y]$ **do** v_y = SECUREAGG($\{v_y^{(i)}\}_{i\in[n]}$)
  Server
    Calculate $v = \sum_x v_x$ and broadcast v, $\{v_x\}$ and $\{v_y\}$ to all the clients.
Round 2: Approximate the statistics
  Server
    Sample the projection matrix $\mathbf{P}$ from $\mathcal{Q}_{2,0,1}^{\ell\times m}$
    Broadcast the projection matrices to the clients
  Client $c_i, i \in [n]$
    Calculate $\bar{v}_{xy} = \frac{v_x v_y}{v}$
    Prepare $\mathbf{u}_i$ s.t. $\mathbf{u}_i[\mathbb{I}(x,y)] = \frac{v_{xy}^{(i)} - \bar{v}_{xy}/n}{\sqrt{\bar{v}_{xy}}}$
    Calculate $\mathbf{e}_i$ = ENCODE(P, u_i)
  $\mathbf{e}$ = SECUREAGG($\{e_i\}_{i\in[n]}$)
  Server
    $\hat{s}_{\chi^2}$ = DECODE(e)

Take a picture to **download** the **full paper**

Experiment



(a) Synthetic Data 1.
(b) Synthetic Data 2.
(c) Synthetic Data 3.
(d) Synthetic Data 4.
(e) Data 1.
(f) Data 2.
(g) Data 3.
(h) Data 4.

Lun Wang, Qi Pang, Shuai Wang, Dawn Song