Privacy Preserving Ridge Regression on Hundreds of Millions of Records

Valeria Nikolaenko, Dan Boneh
Stanford

Udi Weinsberg, Stratis Ioannidis
Marc Joye, Nina Taft
Technicolor
Motivation

Users

Data

Data

Data

Data mining engine

Classification Algorithm, Regression Model, etc.
Motivation

Users

Data

Data

Data

Data mining engine

Classification Algorithm, Regression Model, etc.

Engine learns nothing about data other than final output!
Ridge Regression

Users

\(x_1, y_1\)

\(x_2, y_2\)

\(\vdots\)

\(x_N, y_N\)

\[\beta = \arg \min (\sum (\langle x_i, \beta \rangle - y_i)^2 + \lambda \| \beta \|_2^2)\]

\[= (X^T X + \lambda I)^{-1} X^T y\]

Analyst learns nothing about data other than \(\beta\)

Privacy Preserving Ridge Regression on Hundreds of Millions of Records, Nikolaenko et al.
Our Contributions

- A hybrid architecture
  1. Homomorphic encryption to aggregate user inputs
  2. Yao’s Garbled Circuits to solve linear system

- Implementation
  - Cholesky decomposition implemented as a circuit
  - Regression over $10^8$ users with 20 features in ~10 hours

Privacy Preserving Ridge Regression on Hundreds of Millions of Records, Nikolaenko et al.
### UCI Datasets

<table>
<thead>
<tr>
<th>Name</th>
<th># users</th>
<th># features</th>
<th>Comm. (MB)</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>automobile</td>
<td>205</td>
<td>14</td>
<td>189</td>
<td>100</td>
</tr>
<tr>
<td>communities</td>
<td>1994</td>
<td>20</td>
<td>234</td>
<td>122</td>
</tr>
<tr>
<td>concrete strength</td>
<td>1030</td>
<td>8</td>
<td>27</td>
<td>17</td>
</tr>
<tr>
<td>forest fires</td>
<td>517</td>
<td>12</td>
<td>83</td>
<td>46</td>
</tr>
<tr>
<td>insurance</td>
<td>9822</td>
<td>20</td>
<td>170</td>
<td>92</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- Experiments conducted on 24 real-world datasets at UCI repository.
- 23 out of 24 cases executed in under 150 seconds.
Thank You!