Building and Deploying a Privacy Preserving Data Analysis Platform

Frederick Jansen  ||  fjansen@bu.edu
Director – Software & Application Innovation Lab (sail.bu.edu)
CTO – Nth Party (nthparty.com)
What is MPC?

Where

\[ f(s_1, s_2, s_3, \ldots, s_n) = z \]

private inputs

public output
Privacy-Enhancing Technologies

- Homomorphic Encryption
- Functional Encryption
- Zero-knowledge Proofs
- Differential Privacy
- Multi-Party Computation
MPC for Social Good

- Hey, Charlie
  Combatting Opioid Addiction

- BWWC
  Pay Equity

- Greater Boston Chamber of Commerce
  Economic Inclusion

- The Dataverse Project
  Social Science Research Sharing

- CALLISTO
  Detecting Serial Perpetrators of Sexual Misconduct
Two years ago

- Sugar Beet Auction
  - Aarhus University
  - Danske Sukkerroedyrkere

- Tax Fraud Detection
  - Cybernetica

- Salary Equity
  - BWWC
  - City of Boston

- Corporate Spending on Minority-Owned Businesses
  - Republic of Estonia Tax and Customs Board
  - Greater Boston Chamber of Commerce
Now

MPC Alliance members
Goals:

• Understanding the root causes of the wage gap
• Closing the gap
• Evaluating success

“To make Greater Boston the premier place for working women in America, by closing the wage gap and removing the visible and invisible barriers to women’s advancement.”

BOSTON WOMEN’S WORKFORCE COUNCIL REPORT 2017

Setting the stage
Communication is key

• Should I participate?
  • Access control, encryption at rest, privacy, bad press, trusted 3rd party, cost, expertise, ...

• Can I participate?
  • HIPAA, FOIA, liability, NDA, ...
Additive Secret Sharing #1

Alice

Eve

Business Corp.

Conglomeration Inc.

Average: $8
Clocks for Finite Field Arithmetic

\[ t + r = 9 \]

<table>
<thead>
<tr>
<th>( r )</th>
<th>( \text{Prob.} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>+3</td>
<td>25%</td>
</tr>
<tr>
<td>+6</td>
<td>25%</td>
</tr>
<tr>
<td>+9</td>
<td>25%</td>
</tr>
<tr>
<td>+12</td>
<td>25%</td>
</tr>
</tbody>
</table>

\[ 0 + 9 = 9 \]  \[ 3 + 6 = 9 \]  \[ 6 + 3 = 9 \]  \[ 9 + 12 = 9 \]
Additive Secret Sharing #2

- **Actual**
  - A: $4
  - B: $8
  - C: $3

- **Mask**
  - A: $4
  - B: -$3
  - C: $6

- **Masked Data**
  - Total: $15
  - Average: $5

- **Additive Secret Sharing #2**

- **Total:** $15
- **Average:** $5
Protocol diagram

Contributor A

actual data A + random mask A = masked data A

Contributor B

actual data B + random mask B = masked data B

BU Server (web server/database)

masked data A + masked data B = masked aggregate data

Analyst can never access this data

Public-key Encrypted Storage
only Analyst has key; no one else (including BU) can read the content of this data

Analyst at BWWC (client running web browser)

masked aggregate data

random mask A + random mask B = aggregate mask

actual aggregate data
Other considerations

• Extrapolate from here, trust us
• Build trust through participants and track record
• Other variants of presentations (Shamir’s Secret Sharing) exist
• Average birthday example for in-person settings
• Training sessions and handouts for data submission
• Verbiage? Upload vs submit vs participate?
Why build anew?

- Focus on
- Accessibility
- Familiarity
- Verifiability
- Participation
- Explainability
Version #1

- Developed in 1 week by interns
- Worked, for the most part
Version #3

- Result of extensive user testing
- Supported 100s of participants
- Configurable
- Drag-and-drop

Input your data

Please make sure your session key and participation code match the ones provided in the email sent to you by the Greater Boston Chamber of Commerce. Drag and drop your completed template file to encrypt and include your submission in the aggregate data.

Session key

Participation code

Drag and drop your completed template file here
—or—

Choose file
Verify and submit your data

Please ensure that all data entered is accurate, and confirm that all employees are accounted for by reviewing the total number of employees below.

Totals Check

<table>
<thead>
<tr>
<th>Total Number of Employees</th>
<th>Female</th>
<th>Male</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>15905</td>
<td>16390</td>
<td>32295</td>
</tr>
</tbody>
</table>

Errors

- Invalid session number
- Invalid participation code
- Please answer all Additional Questions

Submission history

- You have not submitted yet

Click the Submit button after reviewing the data for accuracy.
Analytics under MPC

Answer additional questions

We have included these questions to get instant feedback as to how this process went in order to improve the process in future years. Please know that the answers to these questions will be anonymous, and they will be considered separately from the encrypted and aggregated data above.

Which department are you in?
- Human Resources (e.g., HR Manager, HRIS Manager, Compensation Manager, Talent & Development)
- Operations (e.g., Director of Operations)
- Diversity (e.g., Chief Diversity Officer)
- Upper Management (e.g., COO, CEO, Executive Director)
- Other

What kind of HRIS or organizational system does your company/organization use?
- Large-scale traditional HRIS/HRM software (e.g., ADP, Workday, PeopleSoft, etc.)
- Microsoft Office or similar (e.g., Excel, Microsoft Word, Google Docs)
- Other

How easy was it to understand what data was required given the template and instructions?
- Extremely easy
- Moderately easy
- Slightly easy
- Neither easy nor difficult
- Slightly difficult

How long did it take to prepare the data for submission given your organizational system?
- Less than 1 business day
- 1-3 business days
- 4-7 business days
- 7-10 business days
- Greater than 10 business days

Verify and submit your data

Please ensure that all data entered is accurate, and confirm that all employees are accounted for by reviewing the total number of employees below.

Totals Check

<table>
<thead>
<tr>
<th>Employees</th>
<th>Total Number of Employees</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>20000</td>
</tr>
<tr>
<td>Female</td>
<td>10000</td>
</tr>
<tr>
<td>Male</td>
<td>10000</td>
</tr>
</tbody>
</table>

Submission history
- You have not submitted yet

Submit
Quantifying privacy and leakage

• Philosophical question more than technical one
• Some decisions are easy
  • No queries with single row result (maybe)
  • No repeated queries on minor data modifications
• Others not so much
  • How many participants needed?
  • What if participants are anonymous?
  • Which algorithms are appropriate?
• Differential privacy + MPC ideal (sometimes)
All the times we messed up

- Deployment 1: not enough data validation (invalid data)
- Deployment 2: not enough data validation (semantic errors)
- Best effort on browser support, participant asking to fax data
- Assuming wrong upper bound, leaking lower bound
- Recovering from data entry error under MPC, leaking order of magnitude
Acknowledgements

Azer Bestavros  Ran Canetti  Mayank Varia
Nikolaj Volgushev Rawane Issa Kinan Dak Albab
Andrei Lapets Malte Schwarzkopf Lucy Qin
Ben Getchell Assaf Kfouri Shannon Roberts
Ibrahim Shaikh Justin Chen Jacqueline You
Kyle Holzinger Eric Dunton Rose Kelly
Wyatt Howe Ashni Shah Peter Flockhart
Tanner Braun Mike Gajda Mina Michael
Rachel Manzelli Ira Globus-Harris Shreya Pandit

https://multiparty.org
fjansen@bu.edu

CITY of BOSTON
Greater Boston Chamber of Commerce
Massachusetts Bankers Association
Eastern Bank
Boston University
NSF
HRI

Boston University Rafik B. Hariri Institute for Computing and Computational Science & Engineering

Supported in part by NSF Awards #1430145, #1414119, #1718135, and #1739000; supported in part by the IARPA HECTOR program.