ICPSR

SHARING DATA TO ADVANCE SCIENCE

SAFRN: Computing Statistics Without Sharing Private Data

George Alter (University of Michigan, presenter)
Rafail Ostrovsky (UCLA, PI)
Steven Lu (Stealth Software Technologies)
Brett Hemenway Falk (University of Pennsylvania)

SAFRN: Secure Analytics for Reticent Non-consolidated Databases

- Goal: Statistics from multiple private databases computed by Secure Multiparty Computing
- Grant from the Laura and John Arnold Foundation
- Primary grantee: Stealth Software Technologies
  - Rafail Ostrovsky (UCLA) PI
- Subcontract to ICPSR, University of Michigan

How are confidential data shared?

- Anonymization (e.g. HIPAA rules)
- Data use agreements
- Protected environments (physical, virtual)
- Trusted brokers

All of these solutions assume that the data exist in a single place

- Data from multiple sources are together in a combined dataset

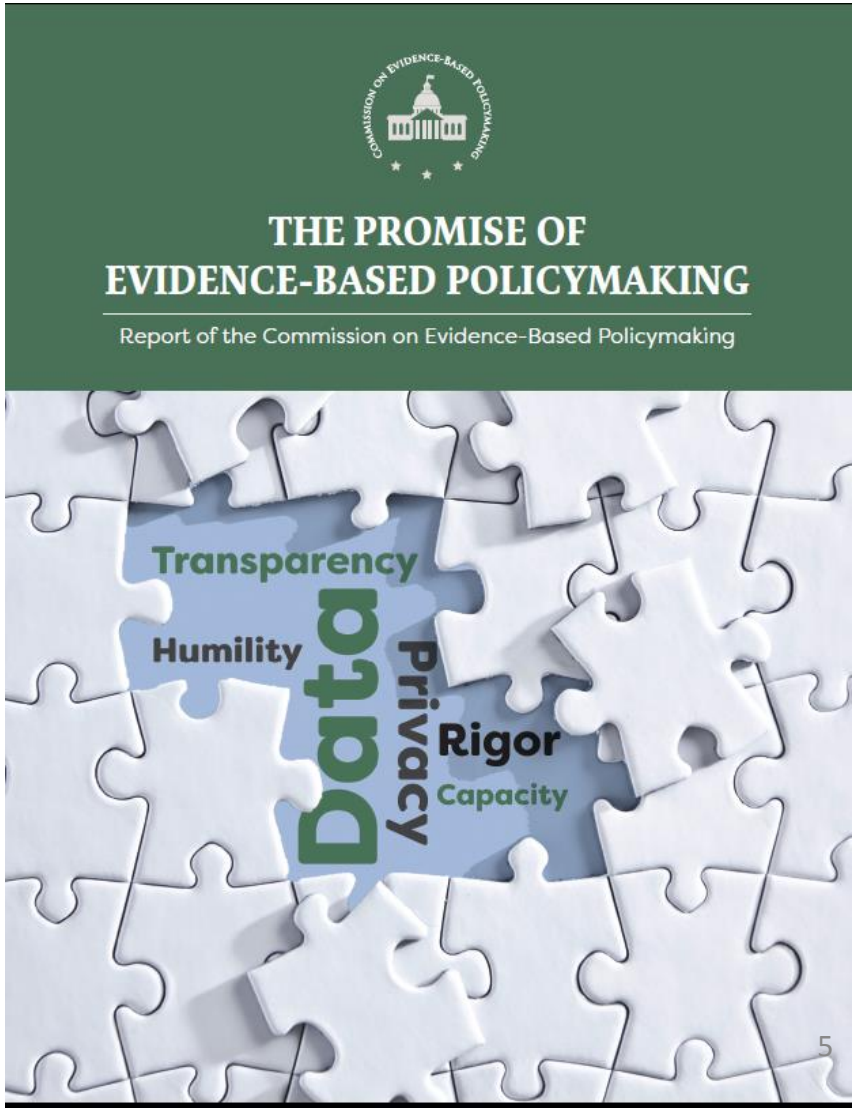
Why use Secure Multi-party Computation?

- Computing summary statistics from multiple confidential sources
- Records must be linked or aggregated across databases
- Databases cannot reveal individual-level (unit record) information to each other or anyone else
- Perform like a trusted broker without ever collecting data in one place

MPC and public policy

"The Commission believes that improved access to data under more privacy-protective conditions can lead to an increase in both the quantity and the quality of evidence to inform important program and policy decisions."

Report of the Commission on Evidence-Based Policymaking, September 2017



MPC appears in proposed legislation

Senator Ron Wyden (Oregon):
"Student Right to Know Before You Go Act"

A BILL

To establish a new higher education data system to allow for more accurate, complete, and secure data on student retention, graduation, and earnings outcomes, at all levels of postsecondary enrollment, and for other purposes.

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

SECTION 1. SHORT TITLE.

This Act may be cited as the "Student Right to Know Before You Go Act of 2017".

(c) CONSIDERATIONS.—In designing, establishing, and maintaining the higher education data system, the Secretary, acting through the Commissioner, shall use the best available cybersecurity and privacy-enhancing technologies to protect the data collected under such system and the privacy of the underlying individuals. In designing the data system, the Commissioner—

(1) shall use secure multiparty computation technologies; or

(2) may utilize technology other than secure multiparty computation technologies if the other technology—

History of MPC

- Theoretical works from the 80s showed that any function can be computed securely (Yao, Goldreich-Micali-Wigderson, Ben Or-Goldwasser-Wigderson, Chaum-Crépeau-Damgård,…)
- Long line of works over the past few decades to make MPC more efficient and usable
- Danish Sugar Beet Auction using MPC
  - It's been 10 years now!
  - Many other examples (Taulbee Salary Survey Attempt, Estonian Financial Data, Boston Wage Study,…)

How does MPC work?

Average Income?

Three people with true salaries S1, S2, S3 they never reveal.
Each computes random numbers Rij to give the other two.
Each shares salary plus Rij given minus Rij received, i.e.,

$$
\begin{array}{rl}
 & X1 = S1 + (R12 + R13) - (R21 + R31) \\
 & X2 = S2 + (R21 + R23) - (R12 + R32) \\
+ & X3 = S3 + (R31 + R32) - (R13 + R23) \\
\hline
\end{array}
$$

$$\text{Sum} = S1 + S2 + S3$$

MPC becoming more widespread: Daniel Goroff from the Alfred P. Sloan Foundation presented this classic example

How does MPC work?

Average Income?

Three people with true salaries S1, S2, S3 they never reveal.
Each computes random numbers Rij to give the other two.
Each shares salary plus Rij given minus Rij received, i.e.,

Only encrypted data are revealed.

$$
\begin{array}{rl}
 & X1 = S1 + (R12 + R13) - (R21 + R31) \\
 & X2 = S2 + (R21 + R23) - (R12 + R32) \\
+ & X3 = S3 + (R31 + R32) - (R13 + R23) \\
\hline
\text{Sum} & = S1 + S2 + S3
\end{array}
$$

MPC becoming more widespread: Daniel Goroff from the Alfred P. Sloan Foundation presented this classic example

How does MPC work?

Average Income?

Three people with true salaries S1, S2, S3 they never reveal.
Each computes random numbers Rij to give the other two.
Each shares salary plus Rij given minus Rij received, i.e.,

Only encrypted data are revealed.

$$
\begin{array}{rl}
 & X1 = S1 + (R12 + R13) - (R21 + R31) \\
 & X2 = S2 + (R21 + R23) - (R12 + R32) \\
+ & X3 = S3 + (R31 + R32) - (R13 + R23) \\
\hline
\text{Sum} & = S1 + S2 + S3
\end{array}
$$

Data owners hold their own encryption keys.

MPC becoming more widespread: Daniel Goroff from the Alfred P. Sloan Foundation presented this classic example

How does MPC work?

- Data owners control their own data
  - Only encrypted data are released
  - The data owner does the encryption
  - The data owner holds the encryption keys
- No one sees unencrypted data
  - Encrypted data are transmitted openly
- Security characteristics of the MPC system can be demonstrated mathematically
- Calculated results are exact
  - Security comes from encryption not from adding noise
- Calculations can be done in real time

What doesn't MPC do?

- MPC outputs may reveal information about an individual
  - Suppose that we compute average income by occupation, age, and state of residence
  - There may be only one US Senator over age 65 from Oregon
- MPC can be combined with other measures (differential privacy) to assure that results do not identify individuals

Limitations of MPC

- MPC cannot be added to existing statistical software
  - MPC algorithms must be engineered into software from the start
- Computations in MPC may be expensive
  - MPC computations produce a large volume of encrypted messages between databases
  - Costs of MPC are difficult to estimate, because they depend on the data

SAFRN Project Goals

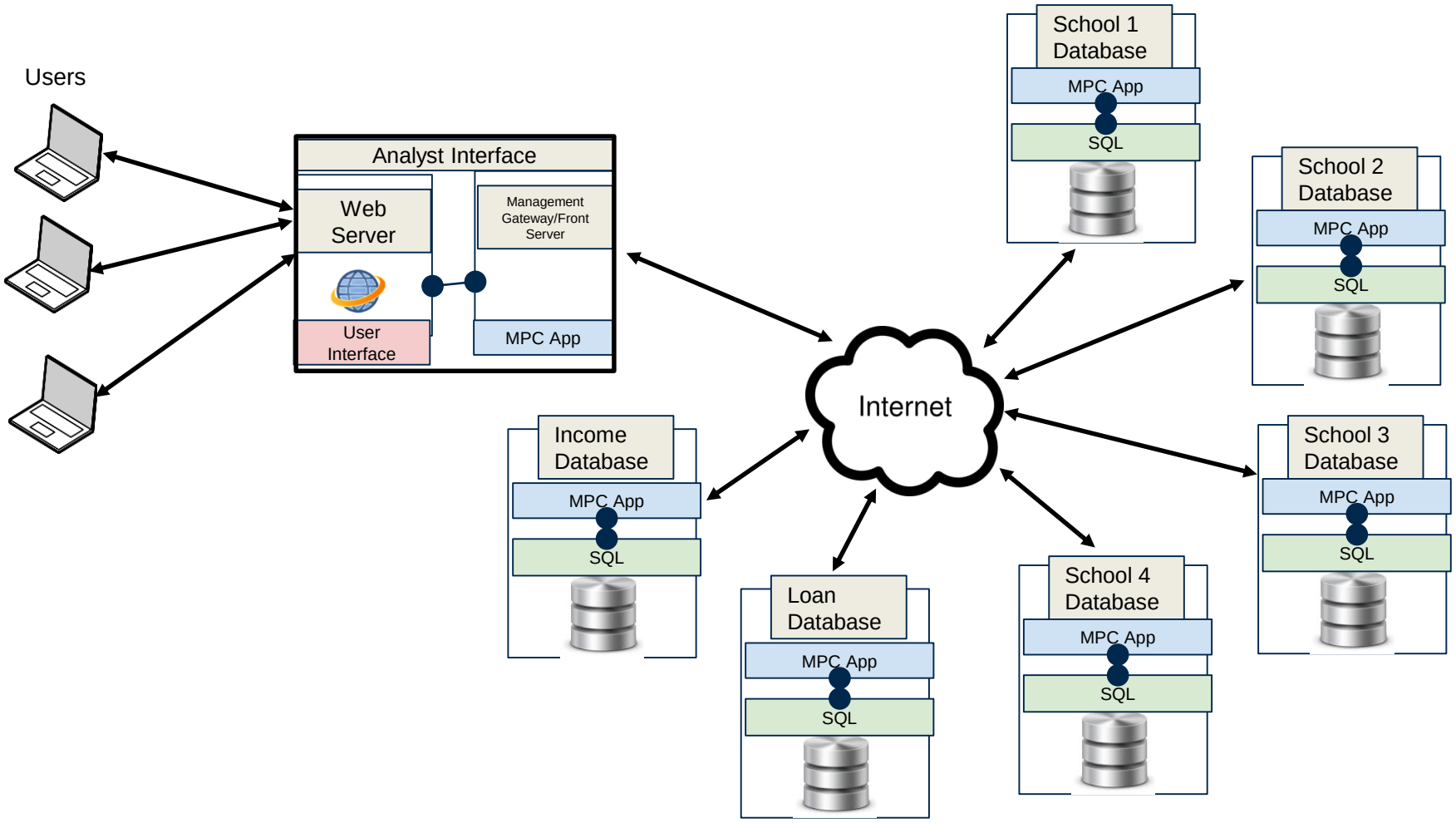
- Demonstrate MPC for computing statistics from multiple private databases
  - Descriptive statistics (e.g. crosstabulation, average)
  - Analytic statistics (e.g. multiple regression)
- Document MPC algorithms
- Estimate costs of MPC using realistic data
  - Synthetic data created from national surveys

SAFRN Prototype

Secure Analytics for Reticent Non-consolidated Databases

- 7 servers
  - 4 Schools
    - 5 Degrees
  - Income server
    - 3 Income variables: 2, 3, and 10 years after graduation
  - Loans
  - Analyst requests statistics
- Computes average incomes and loans by School and Degree
- Synthetic data 1K to 10M cases

SAFRN Prototype



SAFRN Prototype

Secure Analytics For Reticent Non-consolidated Databases (SAFRN)

Statistics computed without sharing private data.

Please select type:

- None
- School
- Degree
- Degree by School

Submit

Show 10 entries Copy CSV Print

Institution	Degree	Average Loan	Average Income 2 years after graduation	Average Income 3 years after graduation	Average Income 10 years after graduation
All	None	5533.16	17591.47	19917.25	35299.26
All	Associate	8880.61	23047.24	25751.07	44224.55
All	Bachelor	23152.17	35054.33	37728.12	54623.43
All	Graduate	82986.23	60455.71	58556.52	78133.03
All	Other	3999.93	20693.20	24658.93	38196.90
All	All	13483.46	24779.19	27165.84	43391.94
Gryffindor	None	5760.65	15757.60	17223.50	33134.19
Gryffindor	Associate	6483.19	27601.43	26123.95	47051.24
Gryffindor	Bachelor	18224.27	38211.45	42058.75	57893.08
Gryffindor	Graduate	70303.50	66160.50	61706.50	112940.00

Funded by laura and john arnold foundation®

SAFRN Prototype

Secure Analytics For Reticent Non-consolidated Databases (SAFRN)

Statistics computed without sharing private data.

Please select type:

- None
- School
- Degree
- Degree by School

Institution	Degree	Average Loan	Average Income 2 years after graduation	Average Income 3 years af
All	None	5533.16	17591.47	19917.25
All	Associate	8880.61	23047.24	25751.07
All	Bachelor	23152.17	35054.33	37728.12

All	Bachelor	23152.17	35054.33	37728.12	54623.43
All	Graduate	82986.23	60455.71	58556.52	78133.03
All	Other	3999.93	20693.20	24658.93	38196.90
All	All	13483.46	24779.19	27165.84	43391.94
Gryffindor	None	5760.65	15757.60	17223.50	33134.19
Gryffindor	Associate	6483.19	27601.43	26123.95	47051.24
Gryffindor	Bachelor	18224.27	38211.45	42058.75	57893.08
Gryffindor	Graduate	70303.50	66160.50	61706.50	112940.00

Funded by laura and john arnold foundation®

SAFRN Team

- Stealth Software Technologies
  - Rafail Ostrovsky (PI)
  - Paul Bunn
  - Brett Hemenway Falk
  - Yuval Ishai
  - Steve Lu
- ICPSR participants
  - George Alter (co-PI)
  - Srujith Cheruku
  - Michael Elliott
  - Stuart Hutchings
  - John Marcotte
  - Dan Pritts
  - Kristine Witkowski

Thank you!

George Alter
altergc@umich.edu

ICPSR

Stealth Software Technologies, Inc.

http://www.stealthsoftwareinc.com/

Funded by

ljaf laura and john arnold foundation®