galois

Inter-Organization Sharing of Sensitive Data for Statistics via Secure Multi-party Computation

David W. Archer, Ph.D., Galois, Inc.
Amy O'Hara, Ph.D., Georgetown University, Massive Data Institute
Rawane Issa, M.S., Galois, Inc.
Stephanie Straus, M.Ed., Georgetown University, Massive Data Institute

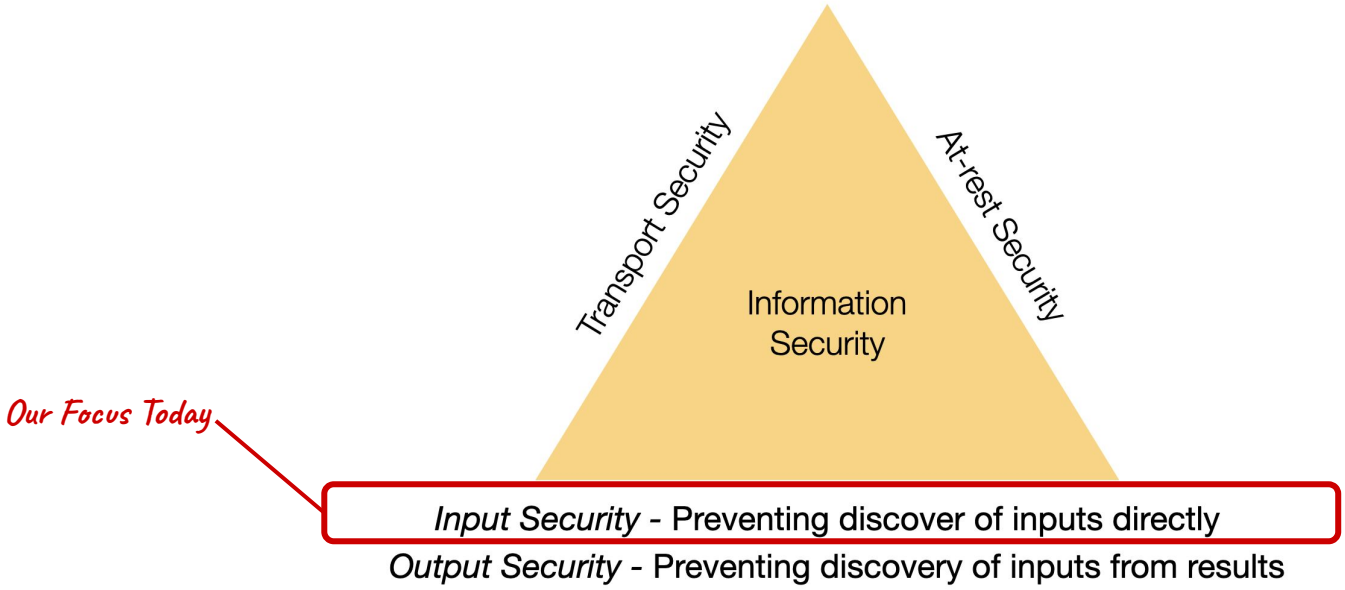
The Utility - Privacy Conundrum

- **Idea**: sharing data makes for better evidence-based policy

- **Opposing idea**: Sensitive data is...umm...sensitive, so sharing should be restricted

- And…current solutions aren't **functioning** well
  - De-identification - Expensive, rarely done right, can harm data utility
  - Secure data centers to host analytics - Doesn't scale
  - Contractual controls - Vulnerable to insider, external threats

The Data Security Domain



Access Control - Preventing leaks of results

One Emerging Solution:

Privacy-Enhancing Technologies

3. Any sufficiently advanced technology is indistinguishable from magic.

Arthur C. Clarke

Ring Of Power Functional Requirements (v1.0)

- *Embody the strength and will to govern*
- *Grant vision from afar*
- *Imbue wisdom and long life*
- *Look good with chain mail*

Ring Of Power Security Requirements

NOT *cause wearer to be in thrall to evil overlord*

(Data Integrity)

NOT *make you reveal secrets to the Armies of Mordor*

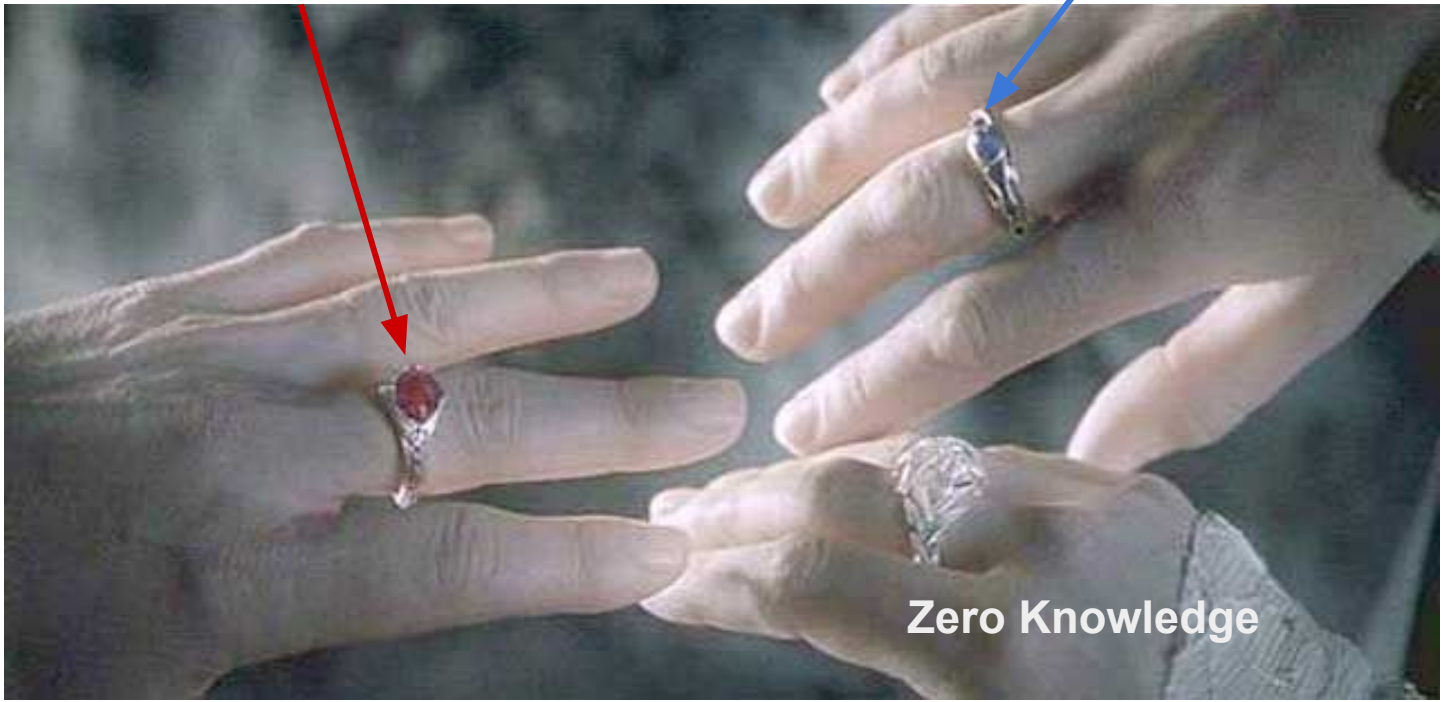
(Data Confidentiality → Privacy)

NOT *cause psychotic episodes of eldritch power*

(we don't really have a name for this yet in Information Security)

Multi-Party Computation (MPC)
Homomorphic Encryption (HE)
Trusted Execution Environments (TEE)

Differential Privacy



Case Study:
Privacy-preserving Statistics
at the
US National Center for Education Statistics

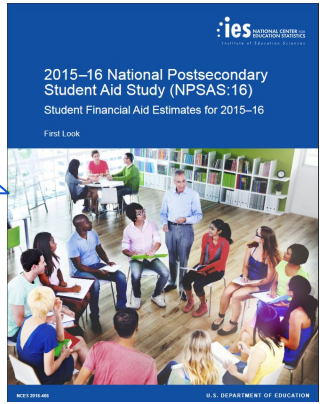
Georgetown University Massive Data Institute and Galois, Inc. thank the Sloan Foundation for their financial support of this project. We appreciate the time and attention from the staff at the National Center for Education Statistics, US Department of Education, for their willingness to partner on the project in pursuit of improved privacy-preserving data-sharing capabilities.

NCES
National Center for Education Statistics

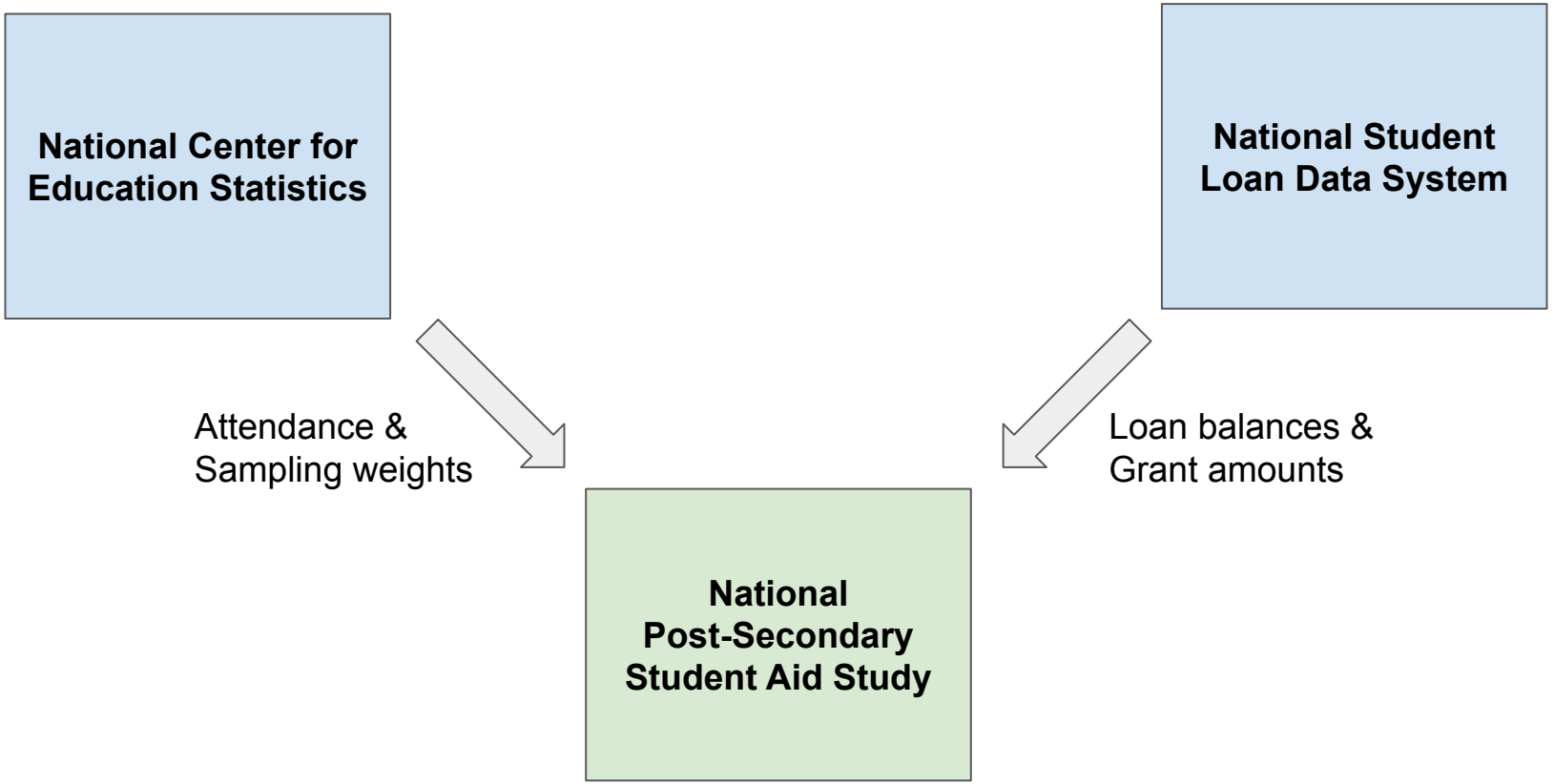
Statistics Computed In Our Initial Experiments

Table 6. Average amounts of federal Title IV aid received by undergraduates from selected programs, by control and level of institution, attendance pattern, dependency status, and income level: 2015–16

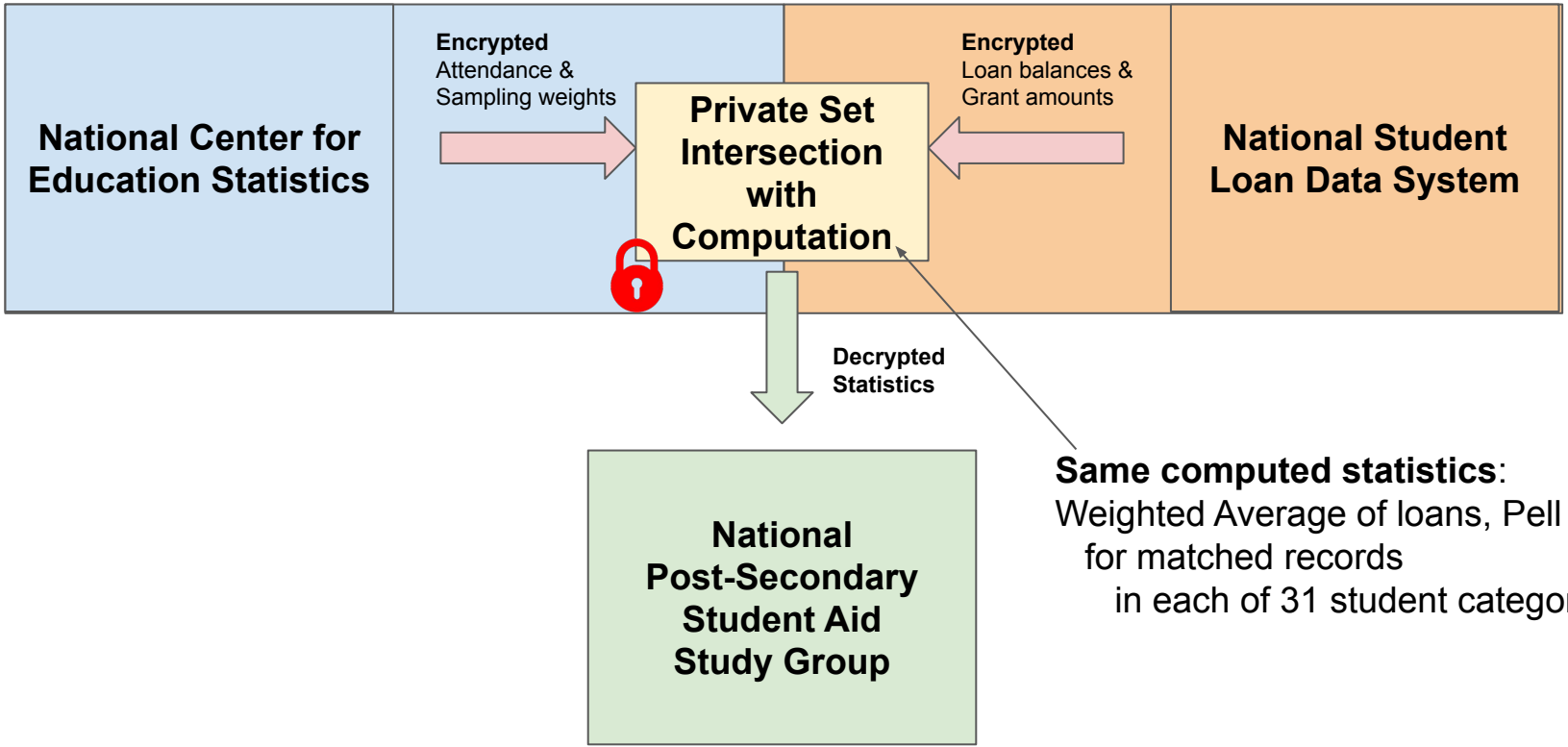
Control and level of institution and student characteristics	Total federal Title IV aid	Federal Pell Grants	Federal campus-based aid[1]	Federal Direct Loans[2]		
				Any	Subsidized	Unsubsidized
Total	**$8,600**	**$3,700**	**$1,700**	**$6,600**	**$3,700**	**$4,000**
All undergraduates						
Control and level of institution						
Public						
Less-than-2-year	5,500	3,300	‡	6,700	3,100	4,500
2-year	4,600	3,300	1,100	4,700	2,900	3,300
4-year	9,400	4,100	1,900	6,600	4,000	4,000
Non-doctorate-granting	7,100	3,800	1,600	6,100	3,700	3,900
Primarily subbaccalaureate[3]	5,100	3,400	1,200	5,200	3,100	3,600
Primarily baccalaureate	8,400	4,000	1,700	6,500	3,900	4,000
Doctorate-granting	10,400	4,200	2,000	6,700	4,000	4,000
Private nonprofit						
Less-than-4-year	9,700	4,100	800	7,000	3,400	4,100
4-year	11,700	4,000	2,500	6,900	4,000	3,900
Non-doctorate-granting	10,900	4,000	2,100	6,800	3,900	4,000
Doctorate-granting	12,300	3,900	2,800	7,000	4,000	3,900
Private for-profit						
Less-than-2-year	8,500	3,700	500	6,400	2,900	3,900
2-year	9,200	3,700	500	7,600	3,500	4,500
4-year	10,900	3,700	800	8,200	3,800	5,000
More than one institution[4]	8,900	3,800	1,600	6,600	3,700	4,100
Attendance pattern						
Full-time/full-year[5]	10,900	4,700	2,100	7,100	4,200	4,100
Part-time or part-year	6,500	3,000	1,100	6,100	3,200	4,000
Full-time/full-year undergraduates[5]						
Dependency and income in 2014[6]						
Dependent students	10,700	4,600	2,300	6,200	4,100	3,400
Less than $20,000	10,800	5,600	2,100	6,100	4,100	2,700
$20,000–39,999	10,900	5,100	2,300	6,200	4,200	2,600
$40,000–59,999	9,900	3,300	2,700	6,200	4,300	2,500
$60,000–79,999	9,500	2,200	2,400	6,300	4,200	2,700
$80,000–99,999	10,600	1,800	2,300	6,300	4,100	3,100
$100,000 or more	11,400	‡	2,200	6,300	4,000	4,600
Independent students	11,400	4,800	1,500	9,400	4,300	6,100
Less than $10,000	12,000	5,300	1,600	9,300	4,300	6,000
$10,000–19,999	11,600	4,600	1,400	9,200	4,300	5,900
$20,000–29,999	11,000	4,200	1,500	9,600	4,500	6,100
$30,000–49,999	10,900	4,900	1,300	9,500	4,400	6,100
$50,000 or more	10,100	3,300	1,900	9,800	4,300	6,800



Current Workflow

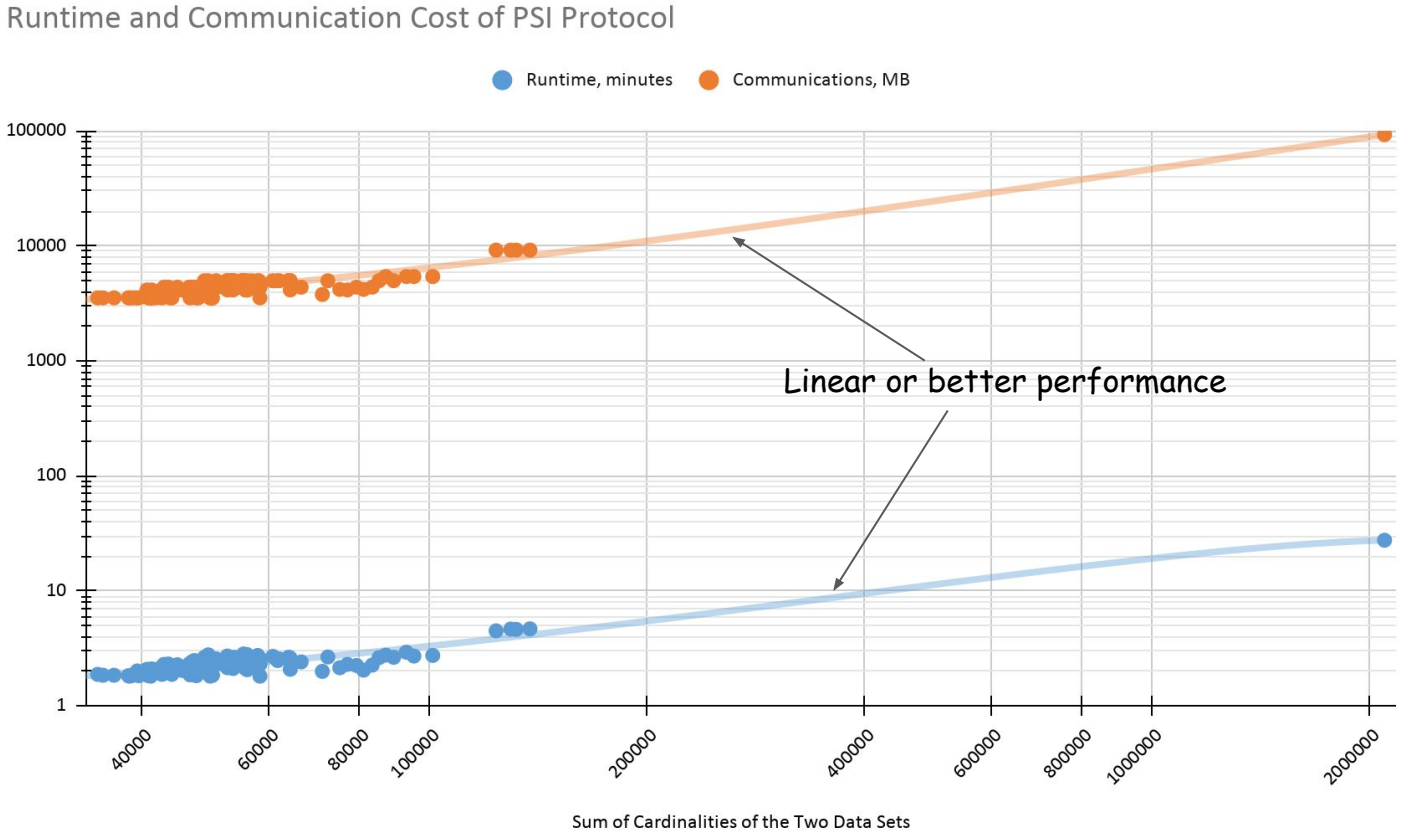


Privacy-Preserving Alternative, via PSI:



Same computed statistics:
Weighted Average of loans, Pell grants
for matched records
in each of 31 student categories

Performance Results: Runtime and Communication Cost vs. Data Set Size



Summary

Successful privacy-preserving, inter-agency data sharing in realistic setting

- **Accurate results**
- **Efficient computation and network costs for real, full-scale data**
- **Scalable - linear at worst in data size**
- **Cost-effective - two cloud instances used for ~5 hours for realistic data in our setting**
- **Privacy-preserving - agencies learn nothing about each others' data**
- **Easy to apply without cryptographic expertise**

| galois |