Sharing Sensitive Department of Education Data Across Organizational Boundaries Using Secure Multiparty Computation

David 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

May 2021

ACKNOWLEDGMENTS: Georgetown University's Massive Data Institute (MDI) thanks the Alfred P. Sloan Foundation for their financial support of this project (GR205852, AWD-7773268). MDI appreciates the time and attention from both the staff at RTI International, for their technical data expertise, and at the National Center for Education Statistics, Department of Education, for their willingness to partner on the project in pursuit of improved privacy-preserving data-sharing capabilities.

Executive Summary

The Federal Data Strategy (FDS) and the Foundations for Evidence-Based Policymaking Act of 2018 (H.R.4174) mandate inter-agency data sharing to promote informed decision making. However, these sharing efforts are often hampered by privacy policy, statute, and practice of agencies. Furthermore, current approaches when such sharing is practiced still put the privacy of shared data at significant risk. Emerging privacy-preserving technologies (PPTs) can enable practical inter-agency data sharing while eliminating this risk to privacy, yet this opportunity has not yet been broadly proven.

To demonstrate the promise of PPTs for this purpose in the US Department of Education, we conducted a pilot project using a cryptographic PPT, secure multiparty computation (MPC), to reproduce a commonplace statistical application that requires sharing of sensitive data among activities within the Department. The prototype application reproduces a portion of the annual 2015–16 National Postsecondary Student Aid Study (NPSAS:16) report, showing statistics on average federal Title IV aid received by undergraduates for the 2015-16 academic year.

In this setting, data comes from two activities within the Department: the National Postsecondary Student Aid Study group (NPSAS) at the National Center for Education Statistics (NCES) and the National Student Loan Data System (NSLDS). Today, preparing the NPSAS reports requires that NSLDS must share sensitive student financial information with NPSAS, and NPSAS must share students' social security numbers with NSLDS. To avoid those disclosures while successfully and efficiently providing the same statistics, our prototype performs the same data linkage and statistical analysis without sharing that sensitive information. Instead, our approach relies on sharing the data and computing the necessary statistics *while the data remains encrypted*. The technique we use -- a variety of MPC called *private set intersection with computation* -- offers cryptographically-proven security at levels comparable to those typically used for encrypting Federal government data today.

The outcome of this project shows that privacy-preserving technologies can be efficiently and effectively applied to sharing of sensitive data among distinct agencies or activities to assure privacy of that data, while also enabling analyses that provide valuable evidence for policy-making. Specifically, our prototype

- produced accurate results for average Federal Pell Grants, Subsidized Federal Direct Loans, Unsubsidized Federal Direct Loans, and all Federal Direct Loans across institution type, attendance pattern, and income level when compared to results from the NPSAS:16 report
- operated efficiently, with computation and network burdens well within practical limits
- demonstrated that users without significant programming experience or cryptographic expertise can use PPTs to protect data privacy in a production-like environment
- and assured that sensitive data used in computing these statistics were kept cryptographically private to the organizations providing that data.

However, PPTs are only useful when put into practice. A well supported introduction of PPTs that addresses certification, legal, technical, and training concerns will be key to successful deployment. In the interim, it would be useful to further demonstrate applicability of PPTs in NCES workflows.

Introduction

In 2017, The Commission on Evidence-Based Policymaking reported that *administrative data* -- data collected during normal operations of the federal government -- should be used to inform evidence-based policy decisions. The Commission unanimously recommended that enhanced inter-agency sharing of administrative data for this purpose should be accompanied by enhanced privacy protections. In so doing, the Commission rejected the notion that such sharing must be concomitant with increased risk of privacy loss. Another of the Commission's recommendations was that the federal government deploy novel privacy-preserving technologies (PPTs) to prevent such privacy loss.

NCES is proactively pursuing the Commission's recommendations. NCES aims to accelerate responsible sharing of Federal data by exploring practical new capabilities that provably protect privacy of sensitive data during computation while preserving utility of the data. Our project is a part of that NCES effort.

In this project, we demonstrate that PPTs such as secure multiparty computation (MPC) are:

- **Relevant to real applications:** We used a specific variety of MPC called *private set intersection with computation* to reproduce a portion of the NCES annual 2015–16 National Postsecondary Student Aid Study (NCES 2018-466, or "NPSAS:16") report. As a result, the parties providing and processing that data learned nothing about data held by each other.
- **Practical and usable:** All computation in the project was performed within the security trust zone of the US Department of Education, using existing resources and support within the IT infrastructure there. All relevant programs were run and managed by our team's domain expert in the application space, who does not have a computer science or programming background.
- **Performant:** We measured computation time and network usage. Our re-creation of this annually-produced result while cryptographically assuring privacy required 4.8 hours of computation time -- minor relative to the data preparation overhead tasks required with or without PPTs. Network usage was also reasonable.
- **Accurate in real world settings:** We compared results from our PPT-based prototype against two separate references: a computation we performed using the standard methodology with the data made available for use in our project; and the statistics published in the NPSAS:16 report. In the former comparison, our results matched the reference results exactly. In the latter comparison, our results matched the reference either exactly or within rounding error[1]. We traced each of the rounding discrepancies to differences between the original data used for the NPSAS:16 report and the data made available for our use[2].

The positive outcome of our project suggests that privacy-preserving techniques are viable for educational statistics at the federal government level. However, legal, regulatory, and other barriers must still be addressed before practical transition to production use is possible. Coordination with the Office of General Counsel (OGC) and the Office of the Chief Information Officer (OCIO) to address legal, regulatory, and

[1] Statistics in the standard methodology used are rounded to the nearest $100.

[2] The source data used in preparation of the NCES 2018-466 report could not be re-created with full accuracy for our use in this project, due to changes in data management practices that occurred in late 2016.

technical barriers will be necessary. While NPSAS is authorized by the Higher Education Act of 1965 (as amended by the Higher Education Opportunity Act of 2008, (20 U.S.C. § 1015(a))) and the Education Sciences Reform Act of 2002 (20 U.S.C. §§ 9541 to 9548), OGC will need to certify that using MPC to compute the statistics for NPSAS does not constitute a disclosure under FERPA (20 U.S.C. 1232g). The OCIO will also need to approve an Authorization to Operate (ATO) for the relevant software packages required, and will need to safelist the relevant application code.

Overview of PPT Used in this Project

In privacy assurance, there are typically three goals:

- **Privacy Goal #1: Input Data Privacy** means that no party involved in computation can access any input data provided by other parties that provide data, nor derive any such input data from intermediate values available during processing, unless the value has been specifically selected in advance for disclosure
- **Privacy Goal #2: Output Data Privacy** means that the output data of the computation do not contain, reveal, or allow derivation of identifiable input data beyond what is specifically allowed in advance by the parties providing the input data
- **Privacy Goal #3: Access Control** means that the application or system includes a mechanism for positive control over which computations can be performed on sensitive input data and which results can be published from those computations

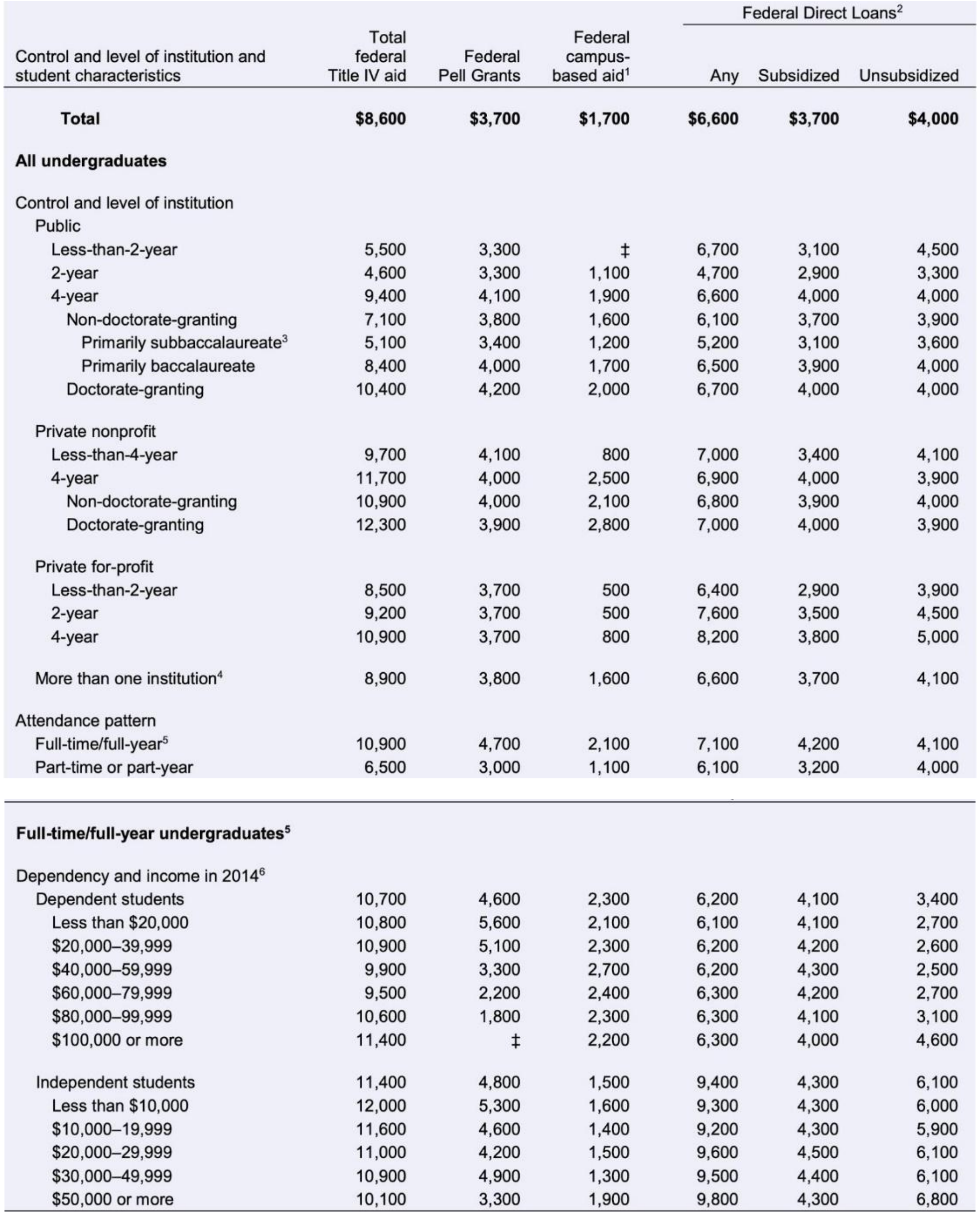
Not all PPTs or alternative information security methods address all of these goals. In this project, we focused on assuring Input Data Privacy because of the nascent technologies available in this area.

The PPT used in this work, *private set intersection with aggregate computation*, is a form of secure multi-party computation (MPC) -- a subfield of cryptography. MPC is compatible with disclosure avoidance techniques that address Output Data Privacy, as well as being compatible with mechanisms typically used for Access Control. MPC deals with the problem of jointly computing an agreed-upon calculation, such as a statistic, among users with limitations that prevent open sharing of data required for that calculation. By employing MPC, calculations can be performed on data *while it remains encrypted*, so that none of those users openly shares their data, and thus none of them (or anyone else) gains information about the input data provided by other users.

An important feature of techniques in the MPC family is that they typically do not require special computation hardware or facilities. MPC algorithms are typically deployed in existing cloud computing installations and require a relatively small number of cooperating computers.

Project Use Case and Scope

The use case for our project is a portion of the periodic NPSAS survey of higher education financial aid. In particular, our project focused on average undergraduate financial aid for a particular academic year, as captured in Table 6 of the 2015–16 National Postsecondary Student Aid Study (NPSAS:16) report. We selected this use case because it is an example of many similar computations used by the Department of Education; and because the scope of this use case made it suitable for a small-scale pilot project.



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

Figure 1. *NPSAS:16 Table 6*. In this project, we re-created all data shown in the Federal Pell Grants column, and all three columns under the 'Federal Direct Loans' heading.

Figure 1 shows NPSAS:16 Table 6, which reports average financial aid amounts, distinguished as Pell grants, unsubsidized student loans, subsidized student loans, and all student loans, across 31 distinct categories of students (shown as separate table rows).

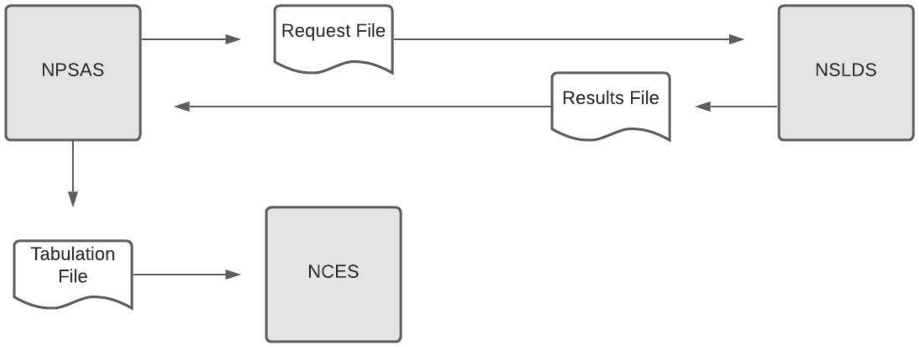


Figure 2. Simplified Diagram of Current (non-Privacy-preserving) Computation.

Today, NCES relies on a trusted third-party contractor to carry out the statistical analysis for the NPSAS reports. As seen in the top of Figure 2, the NPSAS contractor sends a Request File – containing Social Security Numbers (SSNs) of sampled students -- to NSLDS. Using these SSNs, NSLDS queries its grant and loan database, and sends back a Results File that includes any relevant financial aid information for each specified student. Finally, the contractor divides that data into the categories shown as rows in Figure 1, and for each category computes a weighted average, creating the Tabulation File in the bottom left of Figure 2, which is then included in the NPSAS report.

This approach allows NPSAS personnel to view sensitive financial and personal data of students that is private to NSLDS, and allows NSLDS personnel to view the students selected by NPSAS for sampling in the computation. For example, NPSAS learns whether a student either has Pell grants or student loans, and the amounts of those loans and grants across several years of record-keeping. As a result, NPSAS is accountable for assuring the privacy of this data against threats such as data breaches and insider compromise, while NSLDS must fully trust that NPSAS systems are as secure as their own.

In the current approach, NSLDS also learns the social security numbers of students selected for sampling to be included in the NPSAS report. Although perhaps not a direct threat to student privacy (NSLDS already holds the SSNs of all students with loans or grants), this disclosure is an example of another kind of privacy failure: if NPSAS methods in selecting students for sampling were considered sensitive, NSLDS would learn something about the sensitive NPSAS selection methodology by learning the relevant selected SSNs. Thus our use case is an example of a two-way privacy problem.

Our Privacy-Preserving Approach

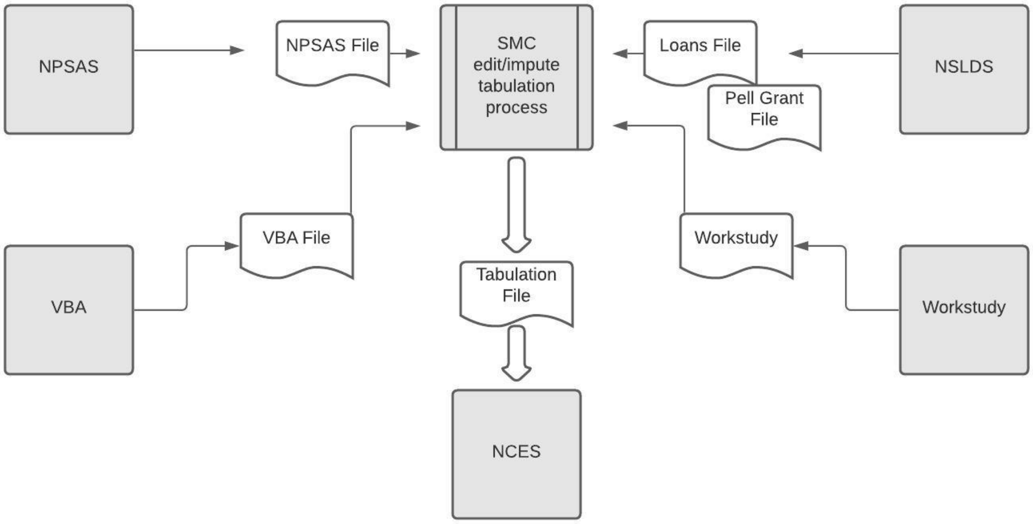
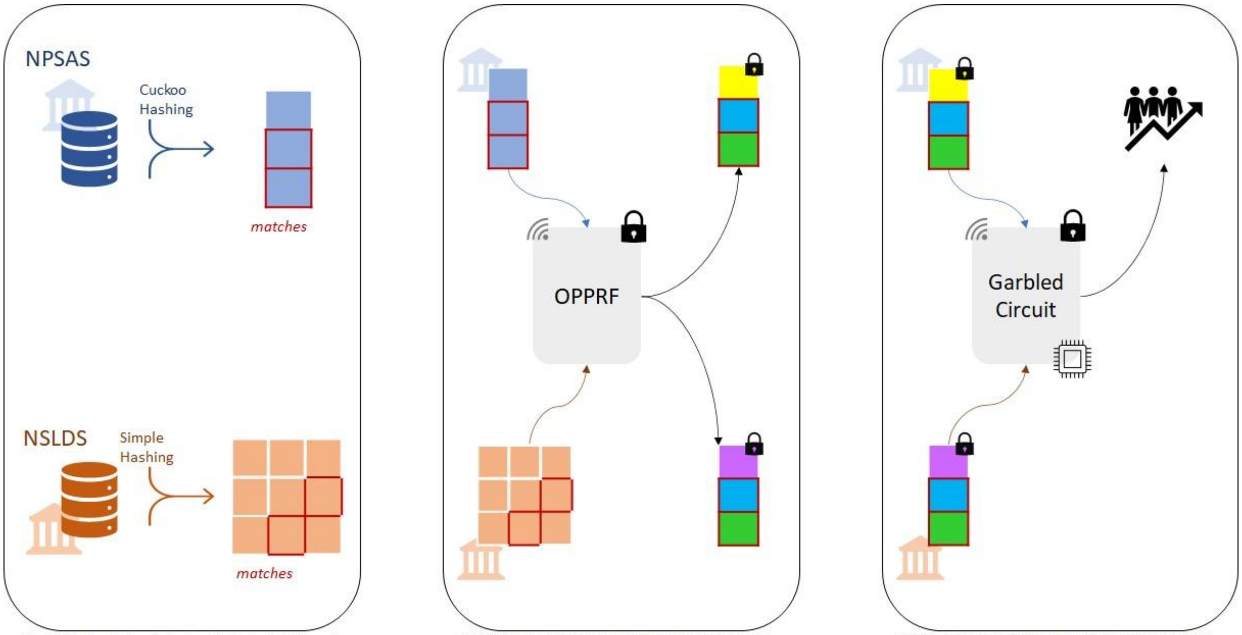


Figure 3. Privacy Preserving Computation Protocol for Results Shown in Figure 1.

In contrast, our project shows that the computations done for Figure 1 can be performed correctly and practically while preventing loss of privacy in both directions. As shown at left in Figure 3, NPSAS personnel select students as before and establish weight factors for the statistical computation, and then provide this data as input to the Private Set Intersection application in encrypted form. In parallel, shown at right in the figure, NSLDS provides the relevant content of their financial aid databases as input to the application, also in encrypted form. The two parties then engage in a shared secure computation (shown at top center), designed using cryptographic techniques that prevent either party from learning anything about the other party's data, except for the statistics output by the application. That output is then provided to NCES (bottom center), who also learns only those statistical results. The MPC protocol we developed and optimized for this project is adapted from a research article published in 2019 [PSTY19]. A more technical description of the protocol we used is below. The reader may skip that description without missing the key top-level messages in this report.

Detailed Protocol Description

The privacy preserving protocol used in our prototype begins at left in Figure 4. Each party -- NPSAS (shown at top in the figure) and NSLDS (shown at bottom) -- work privately on their own data in this protocol step. NPSAS creates a *hash table* and privately uses cuckoo hashing [EMM06] -- a form of hash function collision resolution -- and an agreed-upon set of hash functions to hash the student ID in each selected student record. The resulting hash value is used to select an index, or *bin*, in the hash table.



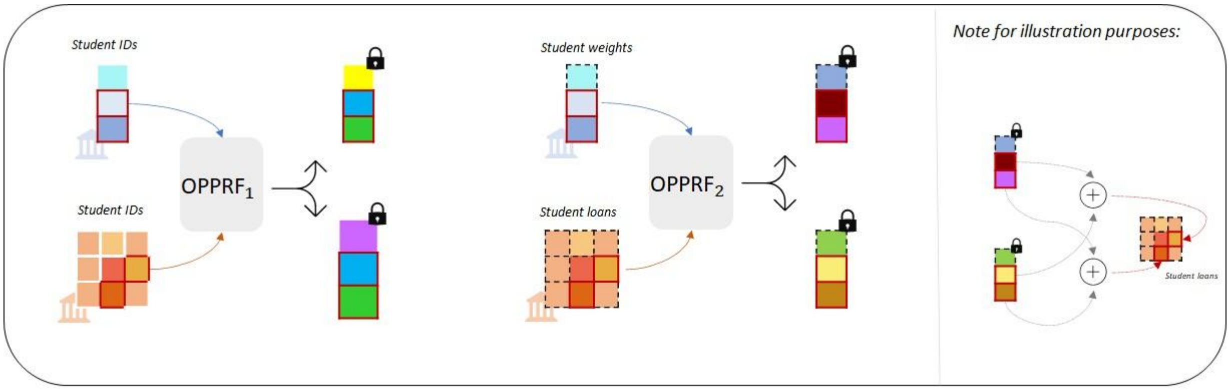
Step 1. Each party locally hashes their student IDs and orders their loans and weights accordingly.

Step 2. Parties jointly invoke OPPRF's for each row in their tables and mask their ID's and values. Matching IDs are masked similarly per bin. Non-Matching IDs are masked randomly. Values follow a similar approach but are not shown for the sake of simplicity.

Step 3. Parties jointly compute a Garbled Circuit over their masked data. Student IDs are securely compared. The values of matching ID's are securely aggregated and weighted.

Figure 4. Overview of the PSI protocol.

The student ID and associated statistical weight are then stored in the selected hash bin. The use of cuckoo hashing assures that there will be at most one entry in any bin in the NPSAS hash table. In parallel, NSLDS creates its own hash table and privately uses the same hash functions used by NPSAS to hash the student ID from each NSLDS student record to select a bin and place the student ID and associated loan or grant values into the appropriate bin in that table. (Note that as a result, an NSLDS bin may contain multiple records). Since both parties agree *a priori* to use the same hash functions, same number of bins, and number their bins in the same order, then a match between a NPSAS record and an NSLDS record results in the matching records appearing in the same bin number in both tables.

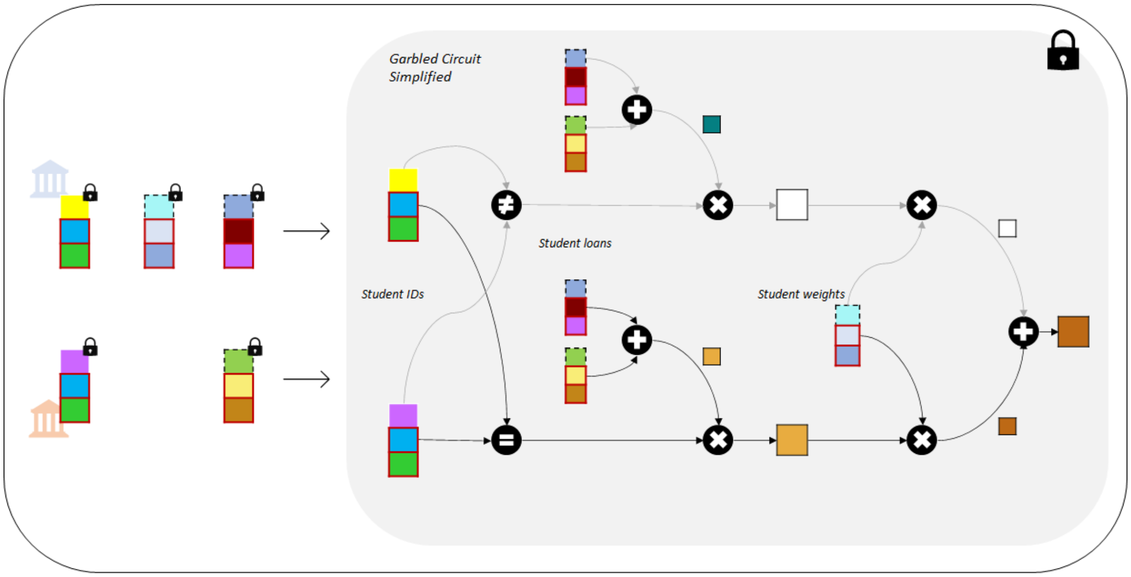


Step 2 (Details). The OPPRF invocations for ID and Loans. Note that in OPPRF2, the sum of the output of the OPPRF for the matching records reconstructs to the original student loan for that record.

Figure 5. Details of the use of OPPRFs shown at center of Figure 4.

The protocol then continues as shown at center in Figure 4 and detailed in Figure 5. In this step, the parties jointly make use of *oblivious, programmable pseudo-random functions* (OPPRFs) [KMP+17]. For each corresponding pair of bins, the NSLDS party first programs an OPPRF with a chosen random value for each ID in her bin. This OPPRF has the property that when the NPSAS party activates the OPPRF and provides the student ID from her corresponding bin as input, she receives as output either the chosen value programmed by NSLDS (if that ID matches her input) or a truly random value (if the IDs do not match). Because both values appear random, the NPSAS party learns nothing about whether there is a match. She retains this returned value and the weight corresponding to the student ID from her bin for processing in the final step of the protocol. Next, the NSLDS party programs a second OPPRF. This OPPRF has the property that when the NPSAS party activates the OPPRF and provides the student ID from her corresponding bin as input, she learns either a masked value of the correct data to include in the statistical computation (if the relevant ID matches her input), or a truly random value (if the IDs do not match), but again cannot learn anything from the returned value. Masking is accomplished here by bitwise addition. She retains this returned value for processing in the next step.

In the final step of the protocol, shown at right in Figure 4 and detailed in Figure 6, NPSAS and NSLDS jointly evaluate a *garbled circuit* (GC) that computes the relevant weighted average. A GC is a cryptographic construction that allows two parties to evaluate a function such that neither party learns the inputs of the other party, nor any intermediate values of the computation. The output of the GC computation is revealed to either or both parties as agreed by them in advance.



Step 3 (Simplified Garbled Circuit). The parties jointly compute a garbled circuit that checks for matching IDs. If IDs match, the circuit successfully reconstructs the student loan, otherwise it computes a random value and nullifies it. In both cases, the values are weighted against the NPSAS weight (in the later case it is again nullified). They are then aggregated. Note that this is a simplified depiction of what happens in the garbled circuit that illustrates the event of matching and unmatching records. In the real circuit, all results are aggregated and normalized.

Figure 6. Details of the garbled circuit computation shown at right in Figure 4.

The garbled circuit we use first checks equality of the student IDs (SSNs) from both parties. If the SSNs match, then the relevant masked input value is unmasked and multiplied by the weight factor, and the result is added to a running total. In the final portion of the garbled circuit, division is performed to convert the running total into an average value for reporting.

As a result of our privacy-preserving protocol, the sample student SSNs from NPSAS are not disclosed to NSLDS; and conversely data about those students are not disclosed to NPSAS -- *not even the data necessary to compute the content of Figure 1*. The only information learned by NPSAS is the weighted average results reported in Figure 1, along with a rough upper bound estimate of how many records might be present in the NSLDS data. Our assumption is that that upper bound is not sensitive information in the context of this data sharing setting.

Experiment Results

We securely computed results for Federal Pell grants, Subsidized Federal Direct Loans, Unsubsidized Federal Direct Loans, and all Federal Direct Loans amounts, in each of the 31 analysis categories shown in Figure 1. Next, we compared our results against a computation we performed using the standard (non-privacy preserving) methodology with the data made available to us for use in our project. Our privacy-preserving computation results matched these reference results exactly. Next, we compared our results against statistics reported in the NPSAS:16 report. In the latter comparison, 86 of 123 results matched the reference exactly, while 37 results of 123 matched within rounding error[3]. Each of the mismatches between our results and the data from the NPSAS:16 report have been traced, with the help of the original data analysts, to a difference in how data was pre-processed in the 2015-16 timeframe compared to pre-processing of the data as we received it. That is, all differences between our results and those shown in the NPSAS:16 report were verifiably attributed to differences in the data provided for our use, and not to differences resulting from our methodology.

Tables 1-4 below show our results compared to the ground truth in the NPSAS report's Table 6. Running times and number of merged records for each computation are shown. Each table shows a different type of federal Title IV aid received by undergraduate students, as shown in the top row of the table.

[3] Statistics in the standard methodology ground truth are rounded to the nearest $100.

	Federal Pell Grant Amounts			
	Ground Truth	Project Results		
	Avg, Grant ($)	Avg. Grant ($)	Runtime (s)	Records Merged
Total	3700	3700	280	39980
Public: Less-than-2-year	3300	3300	125	190
Public: 2-year	3300	3300	129	5130
Public: 4-year	4100	4100	125	8840
Non-doctorate-granting	3800	3800	126	3400
Primarily subbaccalaureate	3400	3400	127	1760
Primarily baccalaureate	4000	4000	127	1640
Doctorate-granting	4200	4200	128	5450
Private nonprofit: Less-than-4-yr	4100	4100	126	570
Private nonprofit: 4-year	4000	4000	127	4550
Non-doctorate-granting	4000	4000	128	2620
Doctorate-granting	3900	3900	135	1930
Private for-profit: Less-than-2-yr	3700	3700	130	1440
Private for-profit: 2-year	3700	3700	122	3740
Private for-profit: 4-year	3700	3700	132	8200
More than one institution	3800	3800	124	7310
Attend. Pattern: Full-time/full-year	4700	4700	138	16940
Attend. Pattern: Pt-time or pt-yr	3000	3000	176	23040
Dependent students	4600	4600	123	8560
Less than $20,000	5600	5600	136	2960
$20,000–39,999	5100	5100	133	3030
$40,000–59,999	3300	3300	130	1890
$60,000–79,999	2200	2200	131	600
$80,000–99,999	1800	1800	130	70
$100,000 or more	Omitted - sample size too small for comparison			
Independent students	4800	4800	129	8380
Less than $10,000	5300	5300	127	3480
$10,000–19,999	4600	4600	121	2180
$20,000–29,999	4200	4200	132	1180
$30,000–49,999	4900	4900	132	1020
$50,000 or more	3300	3300	125	530

Table 1. Average Federal Title IV Pell Grant Results.

	Subsidized Federal Direct Loans Awarded			
	Ground Truth	Project Results		
	Avg. Loans ($)	Avg. Loans ($)	Runtime (s)	Records Merged
Total	3700	3700	270	34110
Public: Less-than-2-year	3100	3100	113	100
Public: 2-year	2900	2800	112	1910
Public: 4-year	4000	4000	109	8250
Non-doctorate-granting	3700	3700	113	2540
Primarily subbaccalaureate	3100	3200	110	1180
Primarily baccalaureate	3900	3900	110	1350
Doctorate-granting	4000	4100	111	5710
Private nonprofit: Less-than-4-yr	3400	3400	111	550
Private nonprofit: 4-year	4000	4000	110	5790
Non-doctorate-granting	3900	4000	109	3090
Doctorate-granting	4000	4100	110	2700
Private for-profit: Less-than-2-yr	2900	2900	111	1100
Private for-profit: 2-year	3500	3500	110	2860
Private for-profit: 4-year	3800	3800	111	7070
More than one institution	3700	3700	112	6490
Attend. Pattern: Full-time/full-year	4200	4200	120	17130
Attend. Pattern: Pt-time or pt-yr	3200	3200	166	16990
Dependent students	4100	4200	129	10150
Less than $20,000	4100	4200	112	1740
$20,000–39,999	4200	4200	111	1930
$40,000–59,999	4300	4300	121	1660
$60,000–79,999	4200	4200	114	1500
$80,000–99,999	4100	4100	109	1180
$100,000 or more	4000	4000	115	2150
Independent students	4300	4300	109	6980
Less than $10,000	4300	4300	112	2470
$10,000–19,999	4300	4300	121	1620
$20,000–29,999	4500	4500	113	1060
$30,000–49,999	4400	4500	119	1050
$50,000 or more	4300	4200	113	780

Table 2. Average Subsidized Federal Direct Loan Results.

	Unsubsidized Federal Direct Loans Awarded			
	Ground Truth	Project Results		
	Avg. Loans ($)	Avg. Loans ($)	Runtime (s)	Records Merged
Total	4000	4100	278	34040
Public: Less-than-2-year	4500	4600	138	100
Public: 2-year	3300	3300	137	1500
Public: 4-year	4000	4000	145	8070
Non-doctorate-granting	3900	4000	145	2330
Primarily subbaccalaureate	3600	3700	145	1050
Primarily baccalaureate	4000	4000	139	1280
Doctorate-granting	4000	4000	138	5740
Private nonprofit: Less-than-4-yr	4100	4200	139	520
Private nonprofit: 4-year	3900	4000	137	5810
Non-doctorate-granting	4000	4000	140	3090
Doctorate-granting	3900	4000	144	2720
Private for-profit: Less-than-2-yr	3900	3900	137	1090
Private for-profit: 2-year	4500	4600	139	2820
Private for-profit: 4-year	5000	4900	139	7360
More than one institution	4100	4200	143	6780
Attend. Pattern: Full-time/full-year	4100	4100	135	17290
Attend. Pattern: Pt-time or pt-yr	4000	4000	163	16760
Dependent students	3400	3400	136	10360
Less than $20,000	2700	2700	136	1330
$20,000–39,999	2600	2700	138	1470
$40,000–59,999	2500	2500	137	1320
$60,000–79,999	2700	2700	149	1300
$80,000–99,999	3100	3100	133	1150
$100,000 or more	4600	4600	144	3840
Independent students	6100	6100	143	6890
Less than $10,000	6000	6000	138	2330
$10,000–19,999	5900	5800	137	1570
$20,000–29,999	6100	6100	135	1030
$30,000–49,999	6100	6100	140	1040
$50,000 or more	6800	6700	140	930

Table 3. Average Unsubsidized Federal Direct Loan Results.

	Any Federal Direct Loans Awarded			
	Ground Truth	**Project Results**		
	Avg. Loans ($)	Avg. Loans ($)	Runtime (s)	Records Merged
Total	6600	6600	281	39920
Public: Less-than-2-year	6700	6900	159	110
Public: 2-year	4700	4700	162	2180
Public: 4-year	6600	6600	160	10160
Non-doctorate-granting	6100	6200	158	3080
Primarily subbaccalaureate	5200	5200	157	1420
Primarily baccalaureate	6500	6500	150	1670
Doctorate-granting	6700	6800	158	7080
Private nonprofit: Less-than-4-yr	7000	7100	167	580
Private nonprofit: 4-year	6900	7000	149	6760
Non-doctorate-granting	6800	6900	157	3610
Doctorate-granting	7000	7100	160	3160
Private for-profit: Less-than-2-yr	6400	6400	154	1170
Private for-profit: 2-year	7600	7600	159	2990
Private for-profit: 4-year	8200	8100	152	8100
More than one institution	6600	6700	149	7880
Attend. Pattern: Full-time/full-year	7100	7100	158	20150
Attend. Pattern: Pt-time or pt-yr	6100	6100	165	19780
Dependent students	6200	6300	159	12460
Less than $20,000	6100	6200	154	1750
$20,000–39,999	6200	6200	156	1960
$40,000–59,999	6200	6300	149	1700
$60,000–79,999	6300	6200	155	1570
$80,000–99,999	6300	6300	163	1340
$100,000 or more	6300	6300	154	4150
Independent students	9400	9400	159	7690
Less than $10,000	9300	9300	156	2650
$10,000–19,999	9200	9100	165	1750
$20,000–29,999	9600	9600	170	1140
$30,000–49,999	9500	9500	168	1160
$50,000 or more	9800	9700	159	990

Table 4. Average Any Federal Direct Loan Results.

Performance

We measured running time and inter-computer communication volume for each of the 123 experiments described in Tables 1-4 above. Because the scale of available data was limited in our use case, we also conducted experiments with substantially larger synthetic data sets (up to 1M records processed).

Our software is written in the Rust language, and is a part of the Galois, Inc. *Swanky* open source secure computation library. Our secure computation application inherently requires two network-connected computers (one held by each participating party). We ran the experiments for Tables 1-4 above on an Amazon EC2 r5.4xlarge instance with 128GB of RAM running Microsoft Windows Server 2016 Datacenter edition, and an Amazon EC2 r5.2xlarge instance with 64GB of RAM running Microsoft Windows Server 2012 edition, in both cases using 4 cores of the underlying Intel Xeon Platinum 8259CL, 2.50GHz CPUs. We ran the synthetic data scalability experiments on two Linux virtual machines, each with 64 GB of RAM and each using 4 of 10 cores on the underlying Intel Xeon Silver 4210, 2.20GHz CPUs.

As shown in Figure 7 and as predicted by an asymptotic performance analysis of our protocol, both the real and synthetic dataset experiment runtimes and communication costs increase proportionally with the size of the two data sets. As the number of records processed approaches 2 million (1M in each set), we observe that runtime for the protocol scales up to roughly 30 minutes, at a communication cost of roughly 100 GBytes.

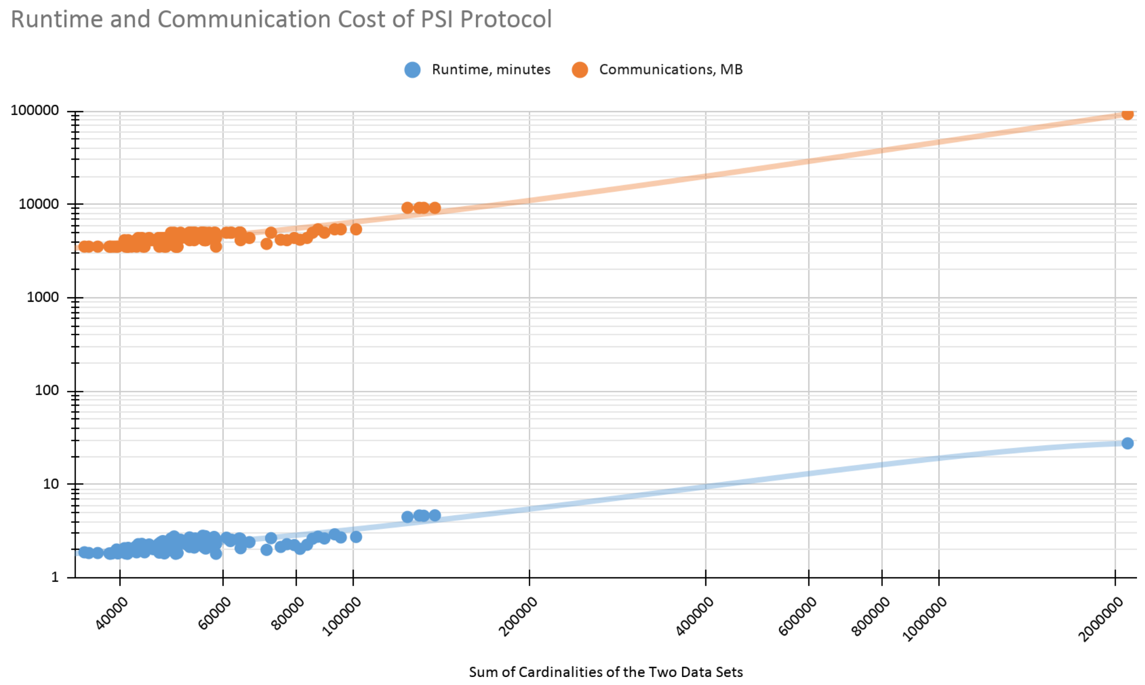


Figure 7. Runtime and Communication Cost Performance as a Function of Records Processed.

Details of Experimental Procedure

The US Department of Education provided an approved member of the project team with secure access to records in the NPSAS Restricted Use File (RUF), which contains de-identified, processed data files for NPSAS and NSLDS, corresponding codebooks, and SAS programs used to derive select source variables.

In the RUF, the NPSAS undergraduate analysis (derived) file, used to create the 31 distinct categories of students corresponding to the 31 Table 6 rows, has been edited and imputed so that there are no missing values, and weighted so that the survey sample is nationally representative of all U.S. postsecondary students.

The NSLDS file in the RUF is not imputed. Instead, it is a universe-level, transactional survey that is housed at Federal Student Aid (FSA) and regularly updated with students' full Pell grants and loans histories. Each record in the NSLDS Loan and Pell file extracts in the RUF corresponds to a grant or loan awarded, so multiple rows in this database can correspond to the same student, and these files must be aggregated accordingly as part of our experiments

Other Related Efforts at PPT Adoption

While acceptance of the promise of PPTs is strong, applications of them have been somewhat limited to date, because 1) agency administrators must first be convinced that PPTs satisfy legal obligations for privacy; 2) adoption must be preceded by *understanding*, which is contingent on a workforce with the ability to educate decision-makers about PPTs and necessary computing infrastructure; and 3) administrators will need to consider how the potential privacy benefits relate to the actual costs of infrastructure needed to deploy PPTs in production.

These reasons notwithstanding, some other agencies have demonstrated PPTs in the service of Input Data Privacy. Secure multi-party computation was shown to assure confidential data sharing among diverse departments within the government of Allegheny County, PA; similar prototypes assured confidentiality of wage data at the Boston Women's Workforce Coalition; and a number of prototype applications showed use of these technologies within the Department of Defense, in particular at DARPA. Alongside these projects, the topic of PPTs has appeared in both legislation and as a topic at a roundtable convened by the White House Office in late 2019.

Alternatives to Input Data PPTs

In this section we briefly describe alternatives to the techniques we used in the project that may also assure Input Data Privacy. Although policy often restricts sensitive data sharing among organizations and government agencies today, some sharing does take place. Current attempts to assure privacy during such sharing generally fall into four categories: (1) de-identification, (2) synthetic substitution, (3) calculations performed by the data provider, and (4) contractual controls. Each has weaknesses that make such sharing either insecure or likely to yield inaccurate results.

Data De-identification

One approach to achieving Data Input Privacy while sharing data is to de-identify the data prior to sharing. De-identification removes or obfuscates input data that might be used to associate data with individuals. De-identification approaches are standard for many federal government agencies, and the practice is encouraged in some federal laws. For example, the Health Insurance Portability and Accountability Act's Privacy Rule requires removal of specific fields such as names, small geographic subdivisions, specific dates, Social Security numbers, and others. Other laws, such as the Confidential Information Protection and Statistical Efficiency Act of 2018, recognize that numerous data fields can be examined in combination, which must be considered when determining appropriate techniques for de-identification and disclosure avoidance.

Unfortunately, de-identification as a form of disclosure avoidance can be expensive, does not offer a verifiable guarantee of confidentiality, requires a high level of technical expertise and precision to be properly implemented, and often restricts data utility, which defeats the purpose of combining source data to produce useful, statistical insights. In addition, de-identification may need to be customized for each new use of data, an expensive effort that must be borne by the owner of the data.

Synthetic Data Substitution

Another approach to Input Data Privacy is to first learn the statistical relationships among the variables in a dataset and then create one or more synthetic datasets that produce approximately the same statistics as those from the original dataset. By creating synthetic data, statistical relationships such as correlations can be approximately replicated for existing data. Unfortunately, synthetic substitution can typically only reproduce known relationships, and only for relatively low-dimensional data (that is, data where records contain no more than 6 or 7 inter-related fields). For example, an expert preparing synthetic substitute data may be aware that the real data show a correlation between two variables, but he or she may not be aware of other correlations in the data - especially higher-order correlations that involve many attribute fields in the data. In many cases, *de novo* research seeks to discover such new relationships, which is impossible when those relationships are lost during the synthesis process simply because they were not already known. In addition, synthetic data construction is an expensive process that also must be borne by the data owner.

Computation by Data Providers

Another approach to Input Data Privacy is for the data provider to perform the computations needed for analytics and pass the results directly to the Output Parties, obviating the need for intermediary Computing Parties. In this setting, Input Parties never need to reveal sensitive, identifiable data and can control which results are released to Output Parties. Input Parties may also bear responsibility for performing risk assessment about disclosure risk avoidance and for performing de-identification activities. Unfortunately, this approach is not scalable because Input Parties must run all analytical queries, regardless of how many Output Parties request multiple analyses. The logistical and computation burden to complete all such requests is untenable for many data providers. Also, this practice suppresses new research on existing data as all computations needed must be conceived of and requested at the time the data is collected. Additional inquiries necessitate additional data collections and associated additional burden.

Policy Implications and Next Steps

As we described above, our pilot project was successful: results were accurate to the ground truth, and computation time was small both in comparison to the repetition rate of the NPSAS report (an annual event) and in comparison to our understanding of the time needed to select students and determine weights prior to the computation. Resource requirements seem practical: the temporary contribution of a computer by each party, and a reasonable amount of network traffic. Usability seems objectively reasonable: experiments were conducted in the usual IT environment within the Department of Education, by a non-expert in privacy preserving technologies.

The success of this project seems to support the notion that privacy preserving technologies such as private set intersection, and MPC more broadly, can (when applied appropriately) assure data privacy while at the same time enabling the combining of data to support evidence-based decision making.

However, we caution that technology is only useful when adopted and put into practice. In this project, we were fortunate to have the support of leadership within the National Center for Education Statistics in the Department of Education to marshal resources within the Department and enable our work. We were also fortunate to have a source of funding to sponsor the project, and the human expertise to put solutions in place. Future prototypes and full deployments will require similar or more extensive support to achieve success. A planned, well supported introduction of technology that addresses certification, legal, technical, and training concerns will be key to successful deployment.

Opportunities for future prototypes are manifold. Projects that create privacy-preserving replacements for data sharing applications across more than just two organizations are one avenue for exploration. Another avenue to explore is the creation of entirely new applications for privacy preserving data sharing -- perhaps applications that to date have been impossible due to the lack of ways to address privacy concerns. We look forward to the opportunity to demonstrate additional capabilities of PPTs in the service of evidence-based decision making.

References

[EMM06] Erlingsson, Ú., M. Manasse and F. McSherry. "A cool and practical alternative to traditional hash tables." In Proceedings of the 7th Workshop on Distributed Data and Structures, pages 1-6, 2006.

[PSTY19] Benny Pinkas, Thomas Schneider, Oleksandr Tkachenko, and Avishay Yanai. Efficient circuit-based PSI with linear communication. In EUROCRYPT, pages 122–153, 2019.

[WSS18] Wine, J., Siegel, P., Stollberg, R. (2018) 2015–16 National Postsecondary Student Aid Study (NCES 2018-466) Data File Documentation (NCES 2018-482). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. Retrieved 11 May 2021 from http://nces.ed.gov/pubsearch.