DE-ID 201 Webinar Meeting Notes

(open to all education privacy working groups and Student Privacy Pledge signatories)

February 12, 2018
Topic: Secure MPC Systems
Hosts: Amelia Vance & Kelsey Finch, Future of Privacy Forum

Moderator: Michael Hawes, Direct of Student Privacy, US Dep't of Education

Boston University Research Team on Basics of Secure Multiparty Computation
Mayank Varia, Research Scientist and Co-Director of RISCS Center
Andre Lapets, Director of Research Development & Institute Fellow
Frederick Jansen, Sr. Software engineer, SAIL Boston University.

1. **Motivation: Calculate Pay Equity**

Speaker: Frederick Jansen, Sr. Software engineer, SAIL Boston University.

Boston: Closing the Wage Gap

- 2013: Boston's Mayor makes pledge to make Boston the best city for working women by closing the gender wage gap.
- The Women's Workforce Council is created, which establishes the "100% Talent" pledge

100% Talent Pledge

- The pledge is signed by over 200 companies
- Signatories pledges 3 things:
  (1) understand root cause of gender wage gap
  (2) use evidence-based practices to close the wage gap
  (3) measuring the progress over time
- Most interested in Goal 3: Evaluating Success: "employers agree to...contribute data to a report compiled by a third party on the Compact's success to date. Employer-level data would not be identified in the report."
- To measure over time, need to collect and measure salary data
    - Friction exists between returning data to a third party while trusting that it won't be identified at an employer-level.

Traditional Approach: Centralize Data Storage & Analysis

- Traditionally, Companies A & B would submit their data to the trusted third party,
- The trusted third party compiles all this data, computes the analytics, and then provides aggregated results to the Boston Women's Workforce Council (BWWC).
- A lot could go wrong:
    - All data contributors must trust third party
    - Data might be vulnerable when it is in transit, stored, or during analysis
    - Third party becomes a target for hackers.
- In the end, lawyers of the third party were not comfortable being liable for storing all this data. If hacked, there's now 200 companies and employees that can sue for breaching their privacy.

- Trivial to reidentify companies, even in larger data set, because they are publicly traded: it is known how many employees they have, how much they spend on salaries, etc.
- Leads to search of a new approach to data storage and analysis [discussed below].

2. MPC for Private Data Analysis

Speaker: Mayank Varia, Research Scientist and Co-Director of RISCS Center

- Cryptography [MPC] enables private data analysis [salaries] for social benefit [pay equity]
- Looking for new workflow without ever revealing any data in the clear

Private Approach: Compute Without Sharing Data

- Underlying idea: Rather than giving data to one trusted third party, companies should provide *something* to multiple computing parties
- Data never leaves your organization in the clear
  - Proactively eliminate risk of data breaches
- Cost of computing increases
- Result revealed to one party
  - Privacy preserved as long as just one computing entity is trustworthy
  - If we didn't want to trust one party, why would we trust two?
    - A: Raw data isn't given to anyone, data is coded and only pieces of information given to each party. Any one of them does not learn anything about the data, but collectively, they can combine to learn something about your data.

Safe Result "In The Open" = Safe Result Under MPC

- Risks of publishing analysis results are same with or without MPC
- Need other privacy tools in order to protect the result
  - De-ID, aggregation, differential privacy
- Benefit of MPC: better accuracy

Private Workflow: Make Decisions Without Sharing

- MPC can be used to make decisions
  - Ex: Is a data analysis across organizations worth it?
    - Can first run MPC to see if answer to joint analysis would even be useful to you before you go through any process of any more sophisticated data sharing
- If analysis is not worthwhile or data is too sensitive, MPC can choose not to compute or reveal the result
- Companies can thus explore opportunities without additional risk

3. Applications of MPC

Speaker: Andre Lapets, Director of Research Development & Institute Fellow

MPC Prototypes & Developments

- 2008: Sugar Beet Auction
- 2010: Financial Data Analysis
- 2013: Satellite Collision
- 2014: Tax Fraud Detection: Republic of Estonia Tax and Customs Board
- 2015–2017: Pay Equity—City of Boston Women's Workforce Council

MPC Enabled for 100% Talent Data Analysis (2015–2017)

- Introduced possibility of using MPC to solve issue of procuring a trusted third party to share data
- Scale: Introduced a web application to 100 different organizations on 3 occasions during 3 MPC sessions to collaboratively analyze and share their information
- For 2 of those analyses a report has been published for 2016 and 2017
- Large sample of the workforce
    - 2016: 69 employers, 113,000 employees, 11% of workforce, $11b in total annual earnings
    - 2017: 114 employers, 167,000 employees, 16% of workforce, $15b in total annual earnings
- Application itself had familiar interface that reflected data that employers were often already collecting internally

100% Talent Compact Data Analysis: Traditional Workflow

- Individual companies could fill out their EEO1C forms and submit them in the clear or through some kind of end-to-end encryption and deliver it to the BWWC
- BWWC could then add up data and get what they wanted, which is the total aggregate spending for each of those categories (level of seniority, gender, ethnicity)
- This was not a workable solution, so we developed a protocol

Private workflow, One Salary at a Time

- Organization can take the true value that they wish to keep private and split it into two values:
    - One is going to be a lie, and they are going to deliver that lie to the other company,
    - the other is going to be the difference between that lie and the true value
- Then, each company can add up the lies and send up the total of the lies to BWWC
- Then, BWWC can add up the total of the discrepancies and subtract them out from the discrepancies, leaving the “true numbers”
- Ran into many issues:
    - Usability—users should not be burdened with understanding how the protocol works or additional responsibilities like storing online keys, so the design was left simple
    - Reviewability—single mistake/incorrect value by a single party may change the entire result. Thus, the client’s end must allow for some kind of feedback for them to check if values entered were incorrect

MPC Technologies: Readiness and Feasibility

- MPC is ready for deployment at least at this scale (small-to-medium data, simple analyses)
- We and others have developed software frameworks for MPC that are stable, scalable, and usable for real deployments
- Today, need domain experts to customize MPC to a scenario
- We are building the automated MPC software of tomorrow

Secure MPC: Summary & Questions

- Secure MPC allows parties to compute over data you cannot see
- Any analysis is possible under MPC, but cost could be higher
- Protects sensitive inputs and intermediate values, but not output
- Independent and complementary to techniques that reduce risk of revealing output
- Fore existing analysis pipelines, eliminate need to centralize trust in a single organization or facility

- For newly desired/required data analyses: eliminate need to create new centralized data warehouses/infrastructures
- You can eliminate the need to centralize trust and data storage within a particular organization

4. Questions

Moderator notes that the majority of questions will be held for the end of the webinar.

You mentioned that any computation is feasible under MPC, how scalable would implementation of MPC framework be for an organization that wants to use this for more sophisticated statistical analysis, and do you need to be able to predict in advance types of analysis to be performed or can you have a framework of MPC that allows for a wide spectrum?

- I'll try to answer all 3 parts:
    - Scalability: It is technically feasible, but it may be expensive to perform computations over large data sets and right now, companies would have to custom-build new applications into their infrastructure that would allow large-scale computation, that's both a software engineering effort, and a deployment effort.
    - Sophistication of the techniques: Today, it is certainly feasible to do things like linear regressions without too much additional cost. To do more sophisticated analyses could be prohibitively expensive today.
    - Need to Know Types of Analyses: Yes, it is a challenge of MPC that you must define your analyses in advance. You are not going to be able to go back through the data and do the analyses you want.
        - Having to decide on the computation in advance could be viewed as a feature rather than bug because calculation must be done in advance, and then people can see it in advance and decide whether it is something they are willing to be a part of, more of a privacy feature.

You mentioned doing more sophisticated work would be more costly, are those costs in the form of customization of the underlying coding, or computing power, or both?

- Currently it's both, you would need to build more custom solutions to scale to larger data sets. You can imagine in 3–4 years there would be a number of vendors on the market that would produce general purpose infrastructures that could be used to enable MPC computations, and you would no longer need to customize, but you would still need to figure out what the analyses you want to run are and whether they have a high communication cost. If there are many parties, that communication overhead could be significant or even prohibitive, even with a turnkey solution.

Efforts to Bring Privacy Enhancing Technologies to Public Policy: A Case Study in Higher Education Data

Speaker: Laura Bernsten, Domestic Policy Advisor, Senate Finance Committee—Minority Staff (Sen. Ron Wyden—OR).

Overview of Higher Ed Data

- Most higher education outcomes are reported through the Integrated Postsecondary Education Data System (IPEDS) right now
- Data is based on surveys of institutions across the country
- Data captures first time, full time students (less than 50% of education)

- As nontraditional students have become the norm, higher education data loses a large swath of the student population, which has implications for data accuracy
- Calculate things like graduation rates, highest credential received, information comparing Pell students as proxy for income metrics, etc.
    - Another thing a better system could do (one that captured individual-level student data), did that student graduate? Did they transfer?
- This matters because consumers care—students and parents want to know whether a degree will deliver what it promises.

2006: Spellings Commission Report

- Called for the creation of "a consumer-friendly information database on higher education with useful, reliable information on institutions, coupled with a search engine to enable students, parents, policymakers, and others to weigh and rank comparative institutional performance."

2008: The Higher Education Opportunity Act's Student Unit Record Ban

- Spellings Commission Report resulted in national-level debate about whether such a database should exist
- Huge concerns over the privacy implications, so the 2008 Higher Education Opportunity Act's reauthorization included a provision expressly prohibiting this type of database.

2012: Student Right to Know Before You Go Act v. 1.0

- When I entered Wyden's office, students were graduating into a pretty terrible recession, and Wyden took a consumer information lesson to this issue and introduced "Know Before You Go 1.0"
- Drafted a bill that essentially linked all state data systems
    - Due to prohibition, many states had created state-level systems
- As a result of that approach, many said "if we don't trust federal govt. why should we trust states? In some ways less efficient, now 50 systems to worry about"
- Data would've included graduation rates, transfer rates by student type, etc.

Student Right to Know Before You Go Act v. 2.0

- The bill did not repeal that prohibition but created an exception for this type of a data system.

2015: Student Right to Know Before You Go Act v. 3.0

- Introduced with many minimal changes

2016: World Changes Regarding Data Privacy

- Started seeing daily headlines about things like the OPM hack, the Equifax breach, and the creation of a Muslim ban and database.
- ACLU declared opposition to the creation of such a student data system
- At the same time, started to hear more about privacy-enhancing technologies such as secure MPC systems (e.g., Data Sugar Beet Auction)
- A commission on evidence-based policymaking was formed by Sen. Murray and Rep. Ryan. One thing that came up over and over again was the student unit record.
    - Given what we were learning, Wyden sent a letter to commission about the new privacy-enhancing technologies the office had been learning about, hoping the commission would consider them.
    - Final report came out in September 2017 and addressed these technologies

2017: Student Right to Know Before You Go Act v. 4.0

- Included explicit mandate that data system be created at Department of Education using privacy-enhancing technologies such as secure MPC.
- The bill didn't explicitly mandate MPC, but outlined what MPC does and said "technologies such as MPC, or at least the same privacy protections"

Conversation is Ongoing At Federal Level

- The healthcare Committee and education and workforce committees having hearings on data issues
- House republicans introduced Higher Education Reauthorization Bill

Questions

Several people, particularly in the higher education community, question whether the technological approach that's included in the latest Wyden bill is actually ready for prime time, would it be better to start with a demo or pilot project?

- Laura Bernsten: When you introduce a bill, you always want to introduce something that's the flashiest possible, why start with a demonstration program if you have a history on introducing legislation calling for the "whole shebang." So, absolutely interested in a demo project and have been having conversations along those lines with several stakeholders. One of the things preventing the data conversation from moving forward is the fact that privacy experts and civil liberties communities have said "there shall not be a data system like that." The data system that a unit record approach would create is at risk of a variety of liabilities. While HEA reauthorization is far off, so now would be a great time to do more demos, there's differing opinions about whether we should have a demo project, while at the same time creating a unit record system not using MPC. I think that approach doesn't give the privacy and civil liberties communities the data security and privacy protection they've been pushing for. In some ways, those approaches are diametrically opposed.

Attendee Questions

Data quality is a big question, MPC is very promising for data sharing particularly between administrative agencies where there are prohibitions against disclosing the underlying PII that has been collected but one of the challenges that agencies, and particularly statistical agencies, face when trying to link with administrative records are the data quality issues and the lack of a true individual identifier. They have to rely on probabilistic linkage and then rely on linkage rates, how viable do you think MPC is when you have to question the underlying certainty of the linkage and you have to provide analysis of your linkage failures?

- BU Team : Whatever process you might use, assuming you did not have anything encumbering you from sharing or viewing all your data in the clear, you can implement those in an automated fashion and run them under MPC. So, if you can come up with a workflow that makes a decision about whether or not the data has sufficient quality or maybe perhaps does transformations to the data based on the quality, or filters out the data, or even does an analysis to determine whether or not the data will be of sufficient quality to make conclusions, all of these things can be, as long as you can automate them and write them as an algorithm, you can run them under MPC and get the result. It is true that it is a challenge to do data harmonization when you don't have access to the actual data. Our own recommendation is that MPC will allow you to ask more questions without revealing that data, which will reduce disclosure and sharing that needs to happen for you to do decide what you want to do and how you want to do it.

The BU team discussed how MPC moves away from trusted third-party approach and instead as long as you can trust one party you're ok, what if you can't trust any?

- BU Team: If you are willing to take on the responsibility of being one of the computing parties, you can take that on yourself, if you don't trust anyone in the set.

we heard from Laura the challenges with education data right now, and we know the BU team has put together some great demonstrations of the potential for MPC, what do you think is a realistic time horizon for implementing MPC as a real and viable solution to the higher education problems Laura discussed?

- Laura Bernsten: It depends on to what extent community embraces this technology—to what extent are resources and attention being poured into this issue. To the extent its resisted it will be slow in adoption. What we here from people in this space is that with appropriate resources and buy-in, it is not far from implementation.
- Andre from BU: Biggest cost of introducing MPC into government infrastructures are actually not associated with MPC but with implementing any new workflow into the process. Other than that we would echo what Laura is saying, a pilot project involves small data is probably possible this year. If we're talking about integrating with existing infrastructures and larger data sets then we're talking a few to several years before that becomes widespread and possible to deploy ubiquitously.

One privacy technology that is frequently discussed as enhancing privacy in the future is synthetic data and simulated data, could MPC be used to create synthetic individual level data that could be provided as public use files to researchers?

- BU Team: Yes. Presumably if you're creating synthetic data, you'd like it to have properties like the actual data, and MPC would allow you to do that. It would involve defining an MPC computation that looks at the real data again under MPC and generates under MPC some kind of synthetic data set that has the same properties.
- Frederic from BU Team: Just wanted to ad that while theoretically possible, there could be a different workflow possible if for example only one organization that holds the data and we want to create a synthetic data set based on that. There are many workflows possible where the algorithm is run behind the firewall and someone would manually inspect that to ensure it's not leaking any information and ship that back to the community. That would not involve the overhead of MPC, MPC is not a one-size-fits-all solution, but should be viewed as one solution of many to solve data privacy issues. If you can find a less complicated solution that works, we would advocate for that.

Would it be possible to implement MPC across a blockchain, and if so, what would that kind of implementation look like?

- BU Team: It is possible, I don't know what such an implementation would look like, but maybe it would help to contrast cryptocurrencies with MPC. Cryptocurrencies are trying to solve similar problem—get an aggregate cohesive picture of the world based on individual inputs and private decision-making. When you build something for a currency though, it has to survive even against competing parties trying actively to thwart it. MPC is trying to solve this problem under not necessarily as strong of a threat model. As a result, MPC does not requiring nearly as massive the computing power as blockchain requires to achieve the consensus mechanisms that it has. By comparison MPC is much simpler, because it doesn't need to address such a threat. If you wanted to store the results of an MPC result on a blockchain you can.

What recommendations would you make to policymakers if they want to move MPC forward?

- Laura Bernsten: One thing that I have learned is that the folks at BU and others are so excited to talk to policymakers about their work. They've been working on MPC for a decade and are proud to talk about it. Thinking about all these thorny policy issues around data versus policy that have been stuck in a sort of standoff will budge and the researchers get very into how a solution would be created to address that thorny policy issue. I would suggest looking at demos that already exist because they're pretty cool.
- Andre from the BU Team: One of the things we experienced when trying to deploy the technology is that we had to spend a lot of time talking to the stakeholders to explain to them what the features and tradeoffs are in using this technology. Once someone sees it actually work they're a lot more interested in implementing it. The most important thing is to enable pilot projects that demonstrate how it may be used.