2017-06-01

Sharing knowledge without sharing data: on the false choice between the privacy and utility...

Sharing Knowledge without Sharing Data

On the false choice between the privacy and utility of information

Azer Bestavros

Computer Science Department
Hariri Institute for Computing
Boston University

June 1, 2017

Some Historical Perspective

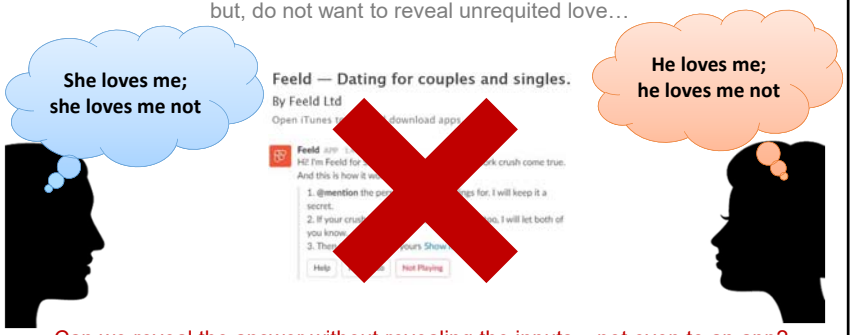
Past: Computers at center of universe
- Computing is centralized & expensive
- There is a dearth of non-synthetic data
- Moving "small data" is cheap
- Best practice: Share data

Today: Data at the center of universe
- There is a deluge of highly-valued data
- Moving "big data" is expensive
- Code is free and computing is cheap
- Question: Why share (private) data?

Azer's Commandment: "Thou shalt not give up/copy/move data, instead move code"

The Valentine Question

Want to know if both parties are interested in each other
but, do not want to reveal unrequited love…



Can we reveal the answer without revealing the inputs – not even to an app?

(Yao's) Millionaires' Problem

Want to know who is wealthier without divulging net worth

Can we reveal the answer without revealing the inputs – not even to an app?

The Dorm Access Question

Want to know if a student is allowed to access the dorm but, do not want to reveal where students go…


The New York Times

N.Y. / REGION

On Campus, a Security Card and More

By AVI SALZMAN OCT. 5, 2003

KEYS have become passé at Quinnipiac University. The private school in Hamden went digital this year, installing hotel-style security locks on doors and allowing students to access everything from dorm rooms to meals using a card-swipe system that records their every move.

"It kinds of bothers me that the university can find out where students go and how long they stay by interrogating locks."

Can we let students in without knowing who they are?

The Labor Department Question

Want to check if Google/Oracle are paying white men more


Department Of Labor

News Release

01/04/2017

US DEPARTMENT OF LABOR SUES GOOGLE INC. FOR COMPENSATION DATA

SAN FRANCISCO – The U.S. Department of Labor has filed a lawsuit to require Google Inc. to provide requested compensation data and documents for the multinational company's Mountain View headquarters as part of a routine compliance evaluation.


USA TODAY 01/18/2017

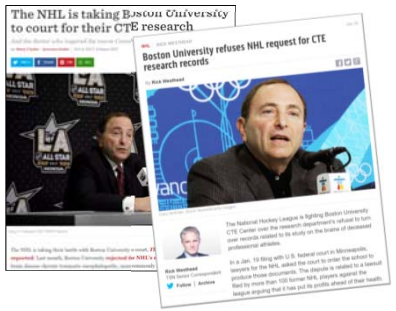
Oracle sued by Labor Department for paying white men more

"In a statement, Google said it balked at turning over the private information of employees."

Can DOL prove (non)compliance without access to sensitive employee records?

The NHL Question

Want to know if players in Chronic Traumatic Encephalopathy study are representative


The NHL is taking Boston University to court for their CTE research

Boston University refuses NHL request for CTE research records

"BU objects to the production of documents concerning the study of the brains of hockey players whose families declined to authorize the release of such information or [those] whose participation was conditioned upon assurances of confidentiality."

BU letter to NHL, 10/26/2015

Can the court get an answer to NHL query without forcing BU to break its promise?

The answer to all these questions is **YES**

We can derive knowledge (K) from data $(x_1, x_2, x_3, \ldots)$ without requiring owners of the data to share it or to trust anything other than mathematics under some assumptions about threats

Azer in the land of social science with mayors, lawyers, CTOs, CIOs, administrators, politicians, journalists, and lawmakers…

A True Story

July 31, 2014


Meeting with Mayor Menino @ BU, July 31, 2014


Katharine Lusk

National Science Foundation
WHERE DISCOVERIES BEGIN

Press Release 14-085

Expanding the breadth and impact of cybersecurity and privacy research

NSF announces two Frontier-scale projects, part of a $74.5 million investment to support foundational cybersecurity research and education

more than 225 projects

UCONN

July 31, 2014

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

FOR IMMEDIATE RELEASE

CONTACT:

(Boston) – Boston University's Rafik B. Hariri Institute for Computing and Computational Science & Engineering today announced it has received funding from the National Science Foundation (NSF) to develop a "smart-city" cloud platform designed to streamline and strengthen multiple municipal functions. Called SCOPE, A Smart-city Cloud-based Open Platform & Eco-system, the project is designed to improve transportation, energy, public safety, asset management, and social services in the City of Boston and across Massachusetts.

BU Initiative on Cities

INITIATIVE ON CITIES

The Initiative on Cities convenes senior urban leaders, academics, and policy makers from around the world to help guide the development of essential services and sustainable infrastructure necessary for cities to flourish.

April 9, 2013

WOMEN'S WORKFORCE COUNCIL

The Women's Workforce Council was established by Mayor Thomas M. Menino on April 9th, 2013— known nationwide as Equal Pay Day. The day marks how far into 2013 women need to work to earn what men earned in 2012. The first of its kind in the country, the Council's mission is to help transform Boston into the best city in the country for working women.

Members of the Council represent the financial, engineering, medical, law, technology and retail sectors, and include small business owners, entrepreneurs, senior executives, as well as academic, labor and nonprofit leaders.

Published October 31, 2013

December 11, 2013

100% TALENT

The Boston Women's Compact

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. By doing so, we will build a more equitable workforce where all talent is cultivated and valued.

GOAL 3

Evaluating Success

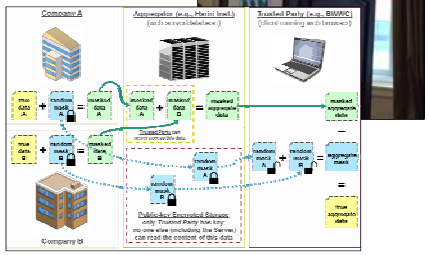
Employers agree to participate in a biennial review to discuss successes and challenges, as well as contribute data to a report compiled by a third-party on the Compact's success to date. Employer-level data would not be identified in the report. The specific data to be reported will build on data already required by federal and state authorities and should not create an additional reporting burden.

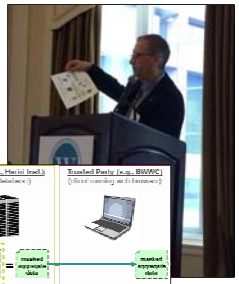
September 4, 2014 ++

A subset of meetings from Azer's Exchange Calendar with BWWC principals, Company CIOs, HR Officers, …

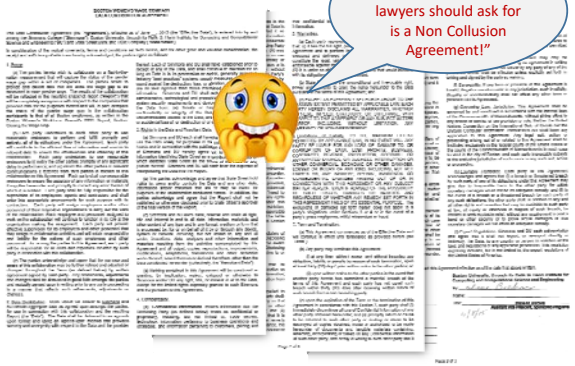
Subject	Start	Duration
Cathy Minehan	Fri 9/5/2014 10:30 AM	2 hours
Simmons College	Mon 10/27/2014 3:30 PM	1.5 hours
Data Collection for Pay Equity	Tue 12/2/2014 11:30 AM	30 minutes
Simmons College people	Fri 1/23/2015 1:00 PM	30 minutes
Invitation: 100% Talent Discussion with Data Partners @ Tue Mar 17, 2015 2pm - 3pm (johnstk3@s...	Tue 3/17/2015 2:00 PM	1 hour
Updated Invitation: MassMutual call with Hariri Institute re: Data Collection... @ Thu May 14, 2015 ...	Thu 5/14/2015 3:00 PM	1 hour
Invitation: Mock collection #1 @ Tue May 19, 2015 11am - 12pm (johnstk3@simmons.edu)	Tue 5/19/2015 11:00 AM	1 hour
Invitation: Mock Collection #2 @ Tue May 26, 2015 11am - 12pm (johnstk3@simmons.edu)	Tue 5/26/2015 11:00 AM	1 hour
Invitation: Mock Collection #3 @ Thu May 28, 2015 11am - 12pm (johnstk3@simmons.edu)	Thu 5/28/2015 11:00 AM	1 hour
Invitation: Call with BWWC @ Wed Jun 3, 2015 11:30am - 12pm (johnstk3@simmons.edu)	Wed 6/3/2015 11:30 AM	30 minutes
Updated Invitation: 100% Talent Data Collection: Hariri and Raytheon @ Fri Jun 5, 2015 9am - 10a...	Fri 6/5/2015 9:00 AM	1 hour
Invitation: 100% TALENT DATA COLLECTION @ Mon Jun 8, 2015 9am - 10:30am (johnstk3@simm...	Mon 6/8/2015 9:00 AM	1.5 hours
Invitation: Meeting with Boston Women's Workforce Council @ Tue Aug 11, 2015 10am - 11am (jo...	Tue 8/11/2015 10:00 AM	1 hour

April 14, 2015

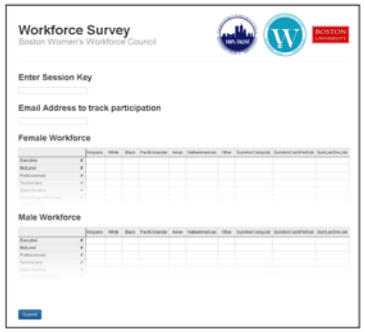




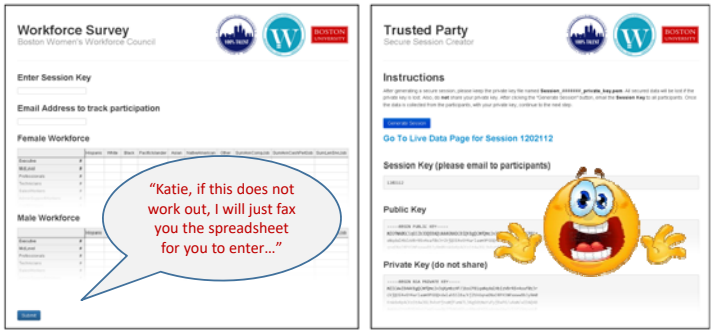
June 6, 2015



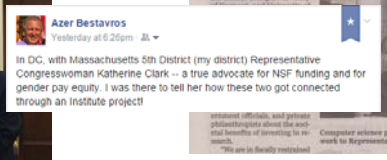
June 8, 2015 (D-day)



June 8, 2015 (D-day)



September 29, 2015



The Boston Globe

The congresswoman, who had signed onto a bill addressing income disparity between men and women, was impressed by the relevance he outlined. *"It's linking it back for the members of Congress,"* Clark said. *"Nobody would think, oh, the Paycheck Fairness Act, how is that tied into NSF funding?"* The meeting was slated for 15 minutes. It lasted 25.

April 28, 2016

The Boston Globe

More Boston businesses join drive to end gender wage gap

100% TALENT
The Boston Women's Compact

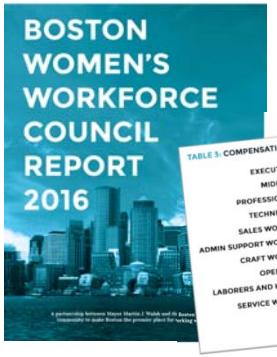
The Sequel

- Compact doubled in size
- More elaborate analytics
- Hardened user interface
- Provide local sanity checks
- Provide comparative metrics

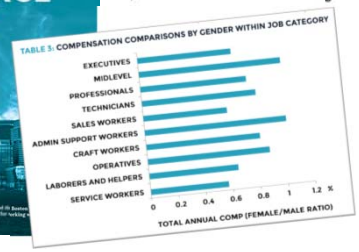
npr wbur ON POINT with TOM ASHBROOK

BWWC co-chair Evelyn Murphy on secure multi-party computation: "*It's used in computer science applications, but it has never been used for public good. Here, we're beginning to show how to use this sophisticated computer science research for public programs.*"

January 5, 2017



"*We collected data regarding 112,600 employees, which represents 11% of the Greater Boston workforce and almost $11 billion in annual earnings.*"



January 27, 2017

Boston Women Workforce Council @ Boston University

CONTACT US:

Boston Women's Workforce Council
Boston University Hariri Institute for Computing
111 Cummington Mall
Boston, MA 02215

Executive Director:

MaryRose Mazzeo

Multi-Party Computation (MPC)

What is it?

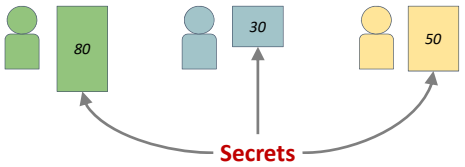
- Given multiple parties p_1, p_2, …, p_n each with private data x_1, x_2, …, x_n
- Parties engage in computing a function $f(x_1, x_2, …, x_n)$
- Nothing is revealing about the inputs beyond what the output of f reveals
- Reasoning about what f leaks is the realm of "differential privacy"

State of the Art

- Theory known since 1979, with Shamir's "How to share a secret"
- Frameworks and libraries increasingly available over the last few years …
- Experience with use cases involving real applications is limited and deployments are not easily portable

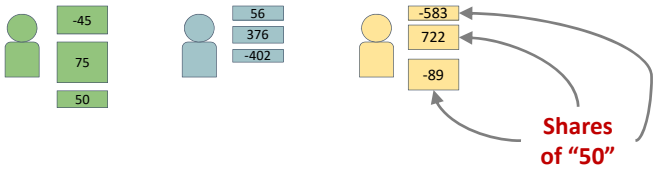
Layman's Example: Sum of Secrets

Can we reveal the sum without divulging the secrets to anybody?

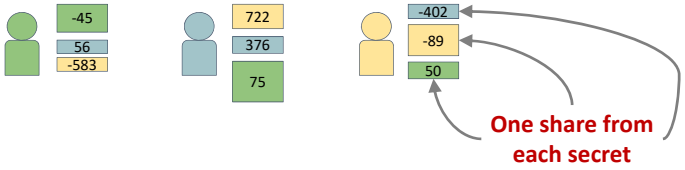


Solution is easy enough to try out with a fourth grader (or lawyers)!

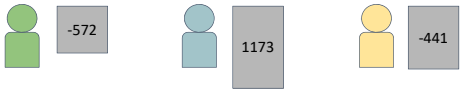
Players split secrets into "shares"



All players exchange shares (securely)



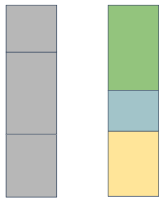
Sum of shares is a share of the sum of secrets!



Exchange/combine to get result

Sum = 160

Lo and behold

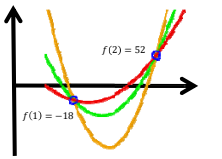


Sum = 160 Sum = 160

Shamir Secret Sharing (1979): The Basic Math

➔ Need $k+1$ points to define polynomial of degree k

- To share a "secret" among k parties, make it the free coefficient of a polynomial $f(x)$ of degree k
- Select coefficients $a_1, a_2, \ldots, a_k$ of $f(x)$ at random
- Give party P_i a "share" of the secret – namely, $f(i)$
- To reconstruct the "secret" all three parties need to combine their shares to find the secret – namely $f(0)$



$$f(x) = 32 - 110x + 60x^2$$

Notes

- Need to use finite field arithmetic to provably avoid any leakage
- Approach allows secret sharing among any number of parties; any subset k can uncover the secret

Multiparty Computing on Secret Shares

Any arbitrary function is a circuit of additions & multiplications

➔ Addition is easy!

- Sum of secrets is represented by $f(x) = f_1(x) + f_2(x)$
- To compute $f(x)$, each party adds its shares of $f_1(x)$ and $f_2(x)$
- Using one round of k messages, sum of secrets can be revealed

➔ Multiplication is not that easy…

- Multiplication of secrets is represented by $f(x) = f_1(x) * f_2(x)$
- Requires $O(k)$ rounds of communications – could be very expensive!

Another Flavor: Yao Garbled Circuits (1986)

- Motivated by Yao's Millionaires Problem (who is wealthier)
- Enables two mistrusting parties to jointly evaluate a function over private inputs using "oblivious transfer" (OT) primitive
  - P_1 replaces inputs of a truth table (gate in circuit) with random labels
  - P_1 encrypts truth table outputs using corresponding input labels
  - P_1 permutes the table and sends the encrypted "garbeled" table to P_2
  - P_1 sends the labels corresponding to its private input to P_2
  - P_1 also sends the labels corresponding to P_2's inputs to P_2 using OT
  - P_2 uses labels corresponding to private inputs to compute output label
  - P_2 communicates output label to P_1 who decrypts it and reveals result

Modeling threats and adversaries

Crypto MPC researchers consider four types of adversaries

- Semi-honest adversary:
  - Follows rules but may attempt to glean information along the way
- Covert adversary:
  - Cheats only if unlikely to be caught
- Rational adversary:
  - Cheats as long as expected payout is larger than expected penalty if caught
- Malicious adversary:
  - Performs any action needed to breach system integrity

The Parties in our MPC Setting

Contributors (100% Talent Companies)

– Have private data needed for computing the analytic
– Number of contributors is unknown in advance

Broker + Analyzer (BWWC)

– Ultimate recipient of the output of the analytic
– May also participate in computing the analytic

Service Provider + Code Distributor (BU)

– Connects/coordinate largely decoupled parties
– Has capacity to (partially) compute the analytic

Threat Modeling & Trust Assumptions

Contributors & analyzers place some trust in each other

– Analyzers trust that contributors will submit valid data
– Contributors trust that analyzers will protect aggregate output
– Contributors trust that analyzers will not collude with others

… but place no trust in service provider

– Service provider cannot be entrusted with data or with the results
– Assume that service provider is incentivized to perform the computation on behalf of the contributors and analyzers

Multi Party Computation: State of the Art

Very active R&D to make MPC accessible to programmers:

Frameworks

- ABY - 2PC with secret sharing and GC; semi-honest adversaries
- batchDualEx - 2PC with GC; malicious adversaries
- Duplo - 2PC GC; malicious adversaries
- Obliv-C - 2PC with gGC; semi-honest adversaries
- Sharemind - 2PC or 3PC with secret sharing; semi-honest adversaries
- SPDZ - General MPC with secret sharing; malicious adversaries
- TinyLEGO - 2PC with GC; malicious adversaries
- Viff - General MPC with secret sharing; semi-honest adversaries

Tools

- CBMC-GC - Creates Boolean circuits (GC) from ANSI-C code
- UC Compiler - Valiant's Universal Circuit Compiler

Primitives

- APRICOT - OT Extension secure against malicious adversaries
- libOTe - Library with various OT Extensions.
- OT Extension - OT Extension secure against malicious adversaries
- SCAPI - Various secure computation API's
- SplitCommit - Additively homomorphic commitment scheme
- TSS - Pure-Rust implementation of threshold secret sharing schemes

Protocols

- BaRK-OPRF - Private Set Intersection
- Linreg - Privacy preserving linear regression
- ORAM (Obliv-C) - Oblivous RAM
- PSI - Private Set Intersection

Commentary on State of Art

Adversarial models are too simplistic

– Need to match crypto threat models with economic, reputation, and legal incentives
– Design of privacy-preserving platforms should take advantage of more realistic models
– Plausible deniability (e.g., participation in MPC) goes beyond keeping data private
– Need to account for the weakest link – the human in the loop!

All parties are not created equal

– Parties may have significantly different backend systems and technical sophistication
– Parties interested in output of MPC may not be the owner of the private data
– Privacy concerns are not uniform across all parties

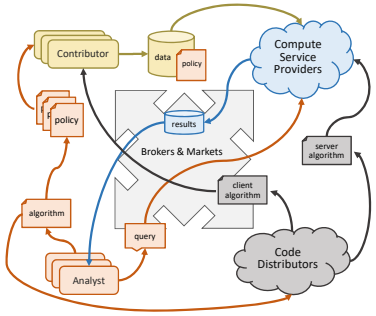
➔ **Need to design solutions that match stakeholders & roles**

Co-Design: Stakeholders & their interfaces

Contributor
- Makes data available
- Sets/chooses policies

Broker or Market
- Identifies data contributors and analysts that have mutually compatible analysis algorithms/data policies
- Helps data contributors craft or choose from possible policies, and helps analysts choose algorithms

Data Analyst
- Defines analysis algorithms
- Chooses performance criteria
- Examines results



Compute Service Provider
- Ingests and stores secret-shared data
- Performs MPC computations
- Optimizes where policies allow

Code Distributor
- Distributes code to contributors and/or service providers
- Provides assurance or auditing that code is from the right analyst

Co-Design Research Agenda: MPC as a Service

Andrei Lapets Kyle Holzinger Eric Dunton Frederick Jansen Nikolaj Volgushev Malte Schwarzkopf Mayank Varia Kinan Bab Rawane Issa

Research Projects @ Boston University

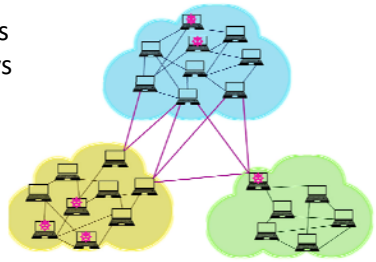
Develop new MPC primitives, toolkits, and optimizations
- Efficient shortest-path algorithms operating over private subgraphs
- Efficient analytics/personalization over private geo-temporal data
- PL and compiler frameworks to expose privacy-utility tradeoffs

Develop MPC "as a service" solutions in various settings
- Web/browser-based MPC as a service platform
- Spark-based MPC platform for Map-Reduce analytics
- Incorporate MPC in big-data cloud workflow management

MPC Primitives for Private Network Analytics

Motivating Scenario: Contagion Risk
- Private (social/computer) networks connected through public gateways
- Some nodes are unsafe, e.g., potentially compromised
- Risk score is shortest distance to an unsafe node

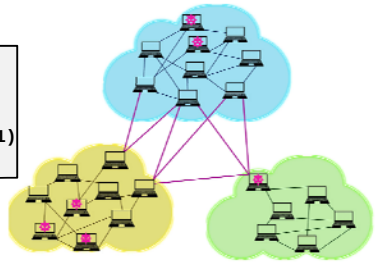
➔ Need to computer risk score



Naïve (not privacy preserving) Solution

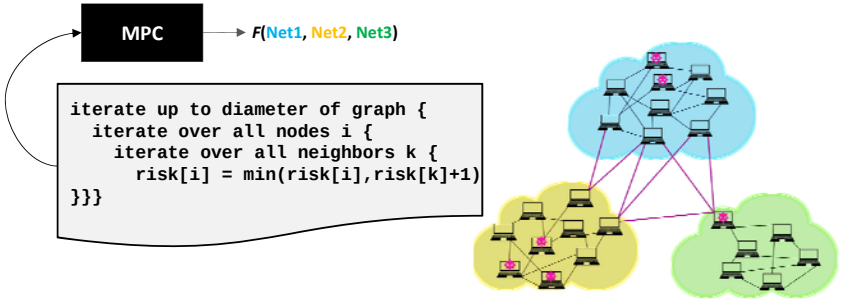
- Assume entire graph is public
- O(diameter * n * degree) algorithm

```
iterate up to diameter of graph {
  iterate over all nodes i {
    iterate over all neighbors k {
      risk[i] = min(risk[i],risk[k]+1)
}}}
```



Can we do this without leaking details about private networks?

Naïve Privacy-Preserving Solution



How efficient would this be?

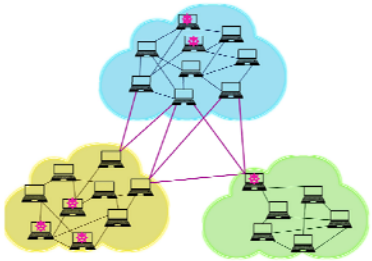
Naïve Privacy-Preserving Solution

P	Node	Edge	MPC[1]	Clear
3	32378	67218	>4hrs	1.7s
3	32	67	4	2s
				2.8s
				2.8s
			Limit	5.8s
5	5	1	Rec. Limit[2]	5.9s
10	108788	250800	-	19.4s

Useless!

[1] Direct implementation using VIFF and python.

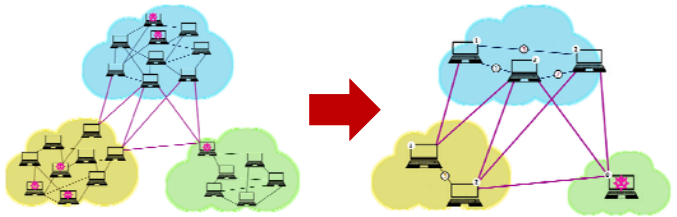
[2] Recursion Limit (10000) was reached.



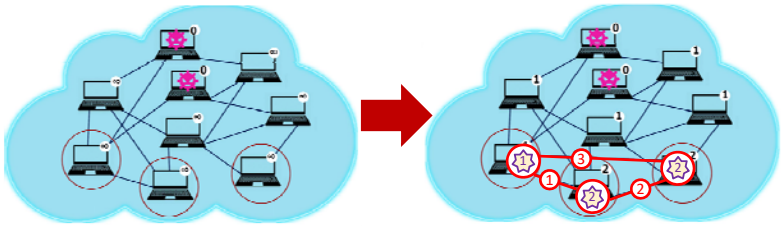
Our Approach: Shrink size of private data!

Key observation #1: For network distance computation, a party needs

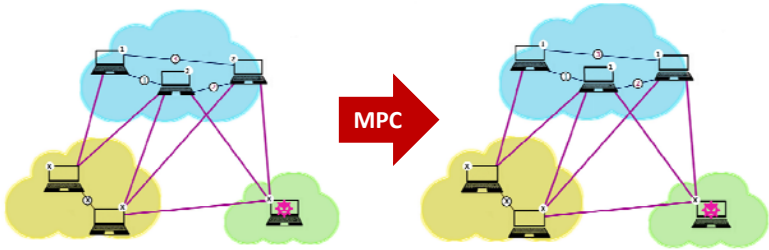
1. The risk score of the gateways of other parties, and
2. A caricature of the topology connecting these gateways



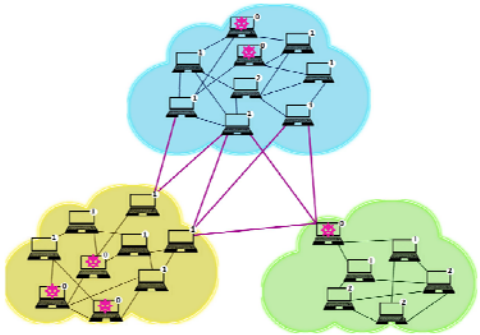
Local Step: Compute Pre-MPC Caricatures



MPC Step: Adjust each party's gateway score

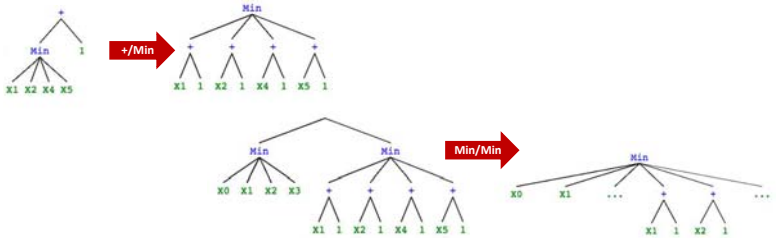


Local Step: Compute final score for all nodes



Our Approach: Optimize MPC stage

Key observation #2: No need to compute all network distance expressions; use semantics of the minimization and addition functions



Our Approach: Implementation & Evaluation

- **Implementation:** in Python using VIFF; design allows for other MPC backends
- **Code Base:** https://github.com/hicsail/ExpressionMPC
- **Experiments:** on peering information from Stanford large network data collection
- **Results:**

P	Node	Edge	Gateway	Pub. Edg.	Our Method	MPC[1]	Clear
3	32378	67218	34	86	0.72min	> 24hrs	1.7s
3	32378	67218	220	579	62min	> 24hrs	2s
4	43510	89783	43	105	2.75min	> 24hrs	2.8s
4	43510	89783	301	850	72min	> 24hrs	2.8s
5	55093	156773	45	105	2min	Rec. Limit[2]	5.8s
5	55093	156773	393	981	154min	Rec. Limit[2]	5.9s
10	108788	250800	44	124	3min	-	19.4s

Research Projects @ Boston University

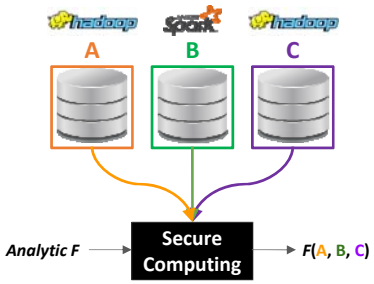
Develop new MPC primitives, toolkits, and optimizations

- Efficient shortest-path algorithms operating over private subgraphs
- Efficient analytics/personalization over private geo-temporal data
- PL and compiler frameworks to expose privacy-utility tradeoffs

Develop MPC "as a service" solutions in various settings

- Web/browser-based MPC as a service platform
- Spark-based MPC platform for Map-Reduce analytics
- Incorporate MPC in big-data cloud workflow management

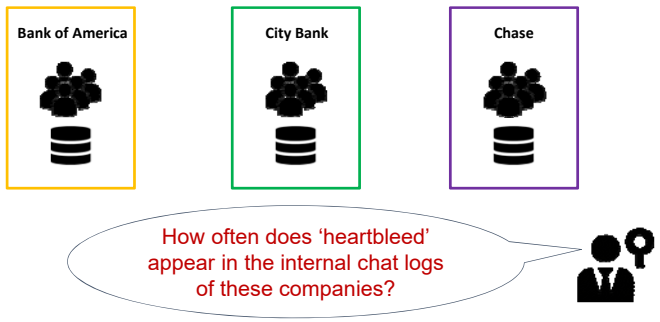
MPC for big-data cloud workflow management



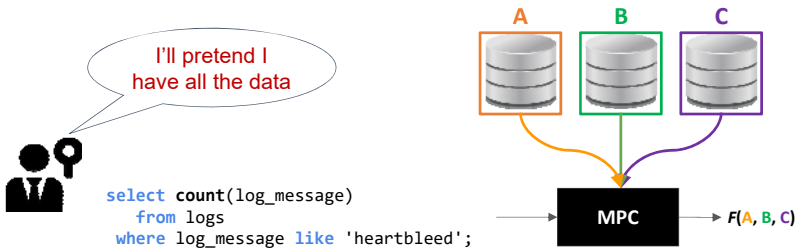
Challenges facing deployments

- **Resources:** MPC can be very slow, resulting in blowup in costs
- **Developers:** MPC frameworks have steep learning curve
- **Info Tech:** Each org works with one data stack, won't use another
- **Info Sec:** Each org works under different data privacy/policy rules

MPC for big-data cloud workflow management



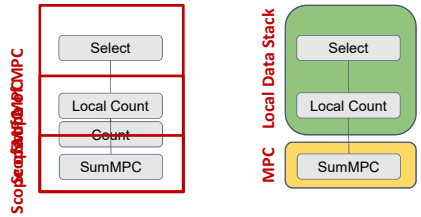
Naïve Approach: MPC on entire function




```
select count(log_message)
   from logs
 where log_message like 'heartbleed';
```

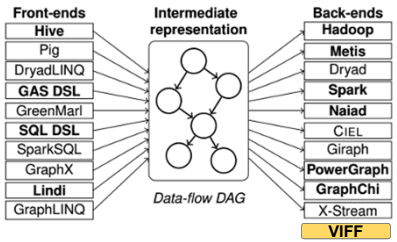

Our Approach: Minimize scope of MPC

```
select count(log_message)
   from logs
 where log_message like 'heartbleed';
```

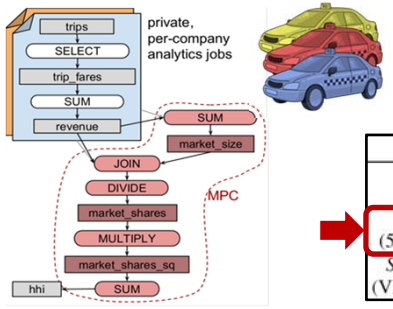


Our Approach: Automate code specialization

Extended Musketeer, a big data workflow manager, to incorporate MPC (VIFF) in backend and to automate code generation



Our Approach: Performance Evaluation



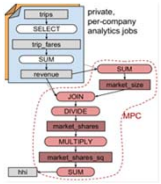
Computing the Herfindahl-Hirschman (market concentration) Index on 156GB of NYC trip data for Taxi companies

Setup	Runtime
Insecure, trusted Hadoop (8 nodes)	16 min 10 s (970s)
Musketeer with MPC (5 parties, 1+1+1+1+4 nodes)	17 min 31 s (1,051s)
Secure MPC framework only (VIFF only, 5 parties, 5 nodes)	>2 hours (7,200s)

MPC for big-data cloud workflow management

Our Solution

- **SQL-like DSL Programming**
  - ➔ *No MPC experience necessary*
  - ➔ *Separate InfoTech from InfoSec*
- **Compiler does MPC transforms**
  - ➔ *No need for privacy experts*
  - ➔ *No need for systems experts*
- **Dispatcher for local deployment**
  - ➔ *No need for new backend*
  - ➔ *No cross-platform integration*



Herfindahl-Hirschman Index on 156GB NYC trip data

Setup	Runtime
Insecure, trusted Hadoop (8 nodes)	16 min 10 s (970s)
Musketeer with MPC (5 parties, 1+1+1+1+4 nodes)	17 min 31 s (1,051s)
Secure MPC framework only (VIFF only, 5 parties, 5 nodes)	>2 hours (7,200s)

MPC as a Service – killer apps…

Systemic Threat Analytics in Federated Settings
- Banking and Finance
- Data Network Operations

Collective Intelligence in Competitive Settings
- Information Brokerage for Business/Marketing Intelligence
- E-Commerce Analytics over Segmented Proprietary Data Assets

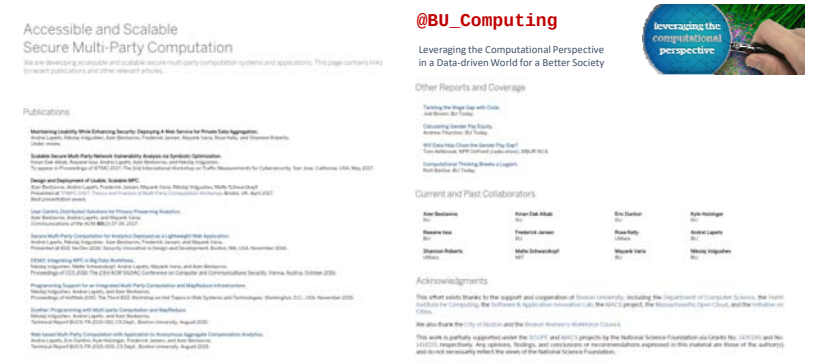
Public Good Analytics
- Anonymous Sensus and Surveys
- Healthcare, Education, and Academic Research
- Compliance Testing/Reporting for Trade Associations

ACSC Advanced Cyber Security Center
MASSACHUSETTS TECHNOLOGY COLLABORATIVE
LexisNexis
BOSTON MEDICAL
MBA Massachusetts Bankers Association

MPC as a Service – it takes a village!

Andrei Lapets Kyle Holzinger Eric Dunton Frederick Jansen Nikolaj Volgushev Malte Schwarzkopf Mayank Varia Kinan Bab Rawane Issa

More information @ www.multiparty.org



Take-Home Message: You can have it both ways

We can derive knowledge (K) from data $(x_1, x_2, x_3, \dots)$ without requiring owners of the data to share it or to trust anything other than mathematics under some assumptions about threats

$$\mathrm{K} = f(\text{CONFIDENTIAL}, \text{CONFIDENTIAL}, \text{CONFIDENTIAL}, \dots)$$

Take-Home Message: Societal Implications

- Privacy/confidentiality concerns should not be used as excuses to deny society the right to answer important questions
- Privacy/confidentiality should not be sacrificed in the name of doing the right thing, or advancing science, or applying the law
- Private data should not be a tradable commodity; computation over private data should be what we offer "for sale"
- Substantial social/financial value can be gained in contexts imposing legal or policy restrictions on sharing raw data

More Perspective about Computer Science

Past: CS is an inward looking
- Study computing fundamentals/limits
- Develop abstractions for programmers
- Make ICT fast, small, cheap, …
- Applied CS seen as engineering

Today: CS is outward looking
- Transforming all academic disciplines
- Developing abstractions for non-CS
- Making society/economy efficient
- Applied CS seen as disrupting society

Azer's Perspective: "Applied CS is pivoting from engineering to social sciences"