SQL over Private Data

Ke Yi
CSE, HKUST

Yao's Millionaire Problem (1982)

x y

Want to know if $x > y$ without revealing private data

Solution: Yao's garbled circuit

SQL over Private Data

- Example

  Insurance Company: R_1(person,coinsurance,state) and R_3(disease,class)

  Hospital: R_2(person,disease,cost)

  The insurance company wishes to estimate the amount of payment it would pay out, classified by disease types, before the patients submit claims.

  This can be expressed by the following query:

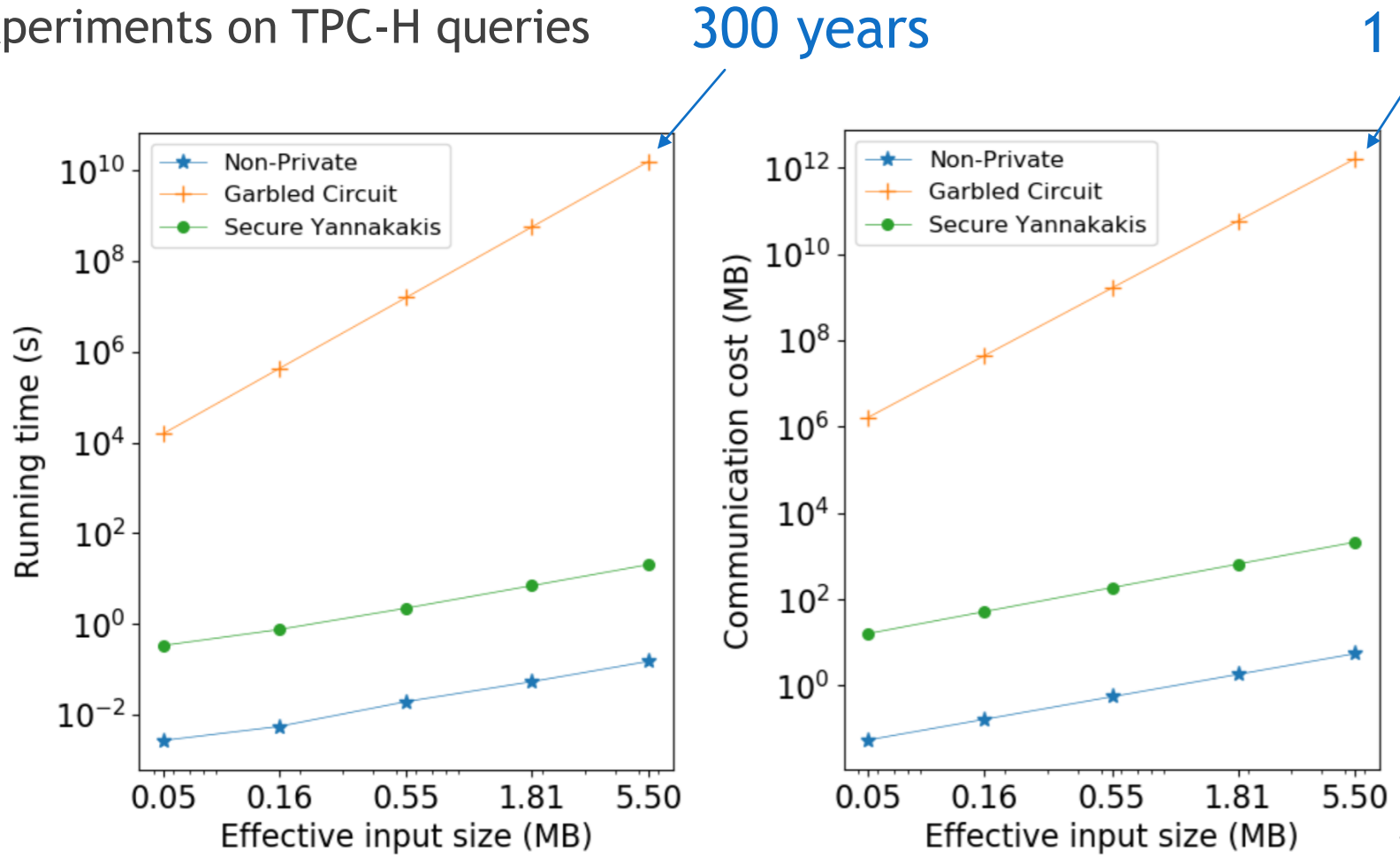
```
select class, sum(cost * (1 - coinsurance))
from R1, R2, R3
where R1.person=R2.person and R2.disease=R3.disease
group by class;
```

  Security guarantee in the multi-party computation (MPC) model:

  - The insurance company can learn the query results, but nothing else about the hospital's data.
  - The hospital cannot learn anything about the insurance company's data.

Our Results (SIGMOD 2021)

- Experiments on TPC-H queries



What if they want to know $x + y$?

x y

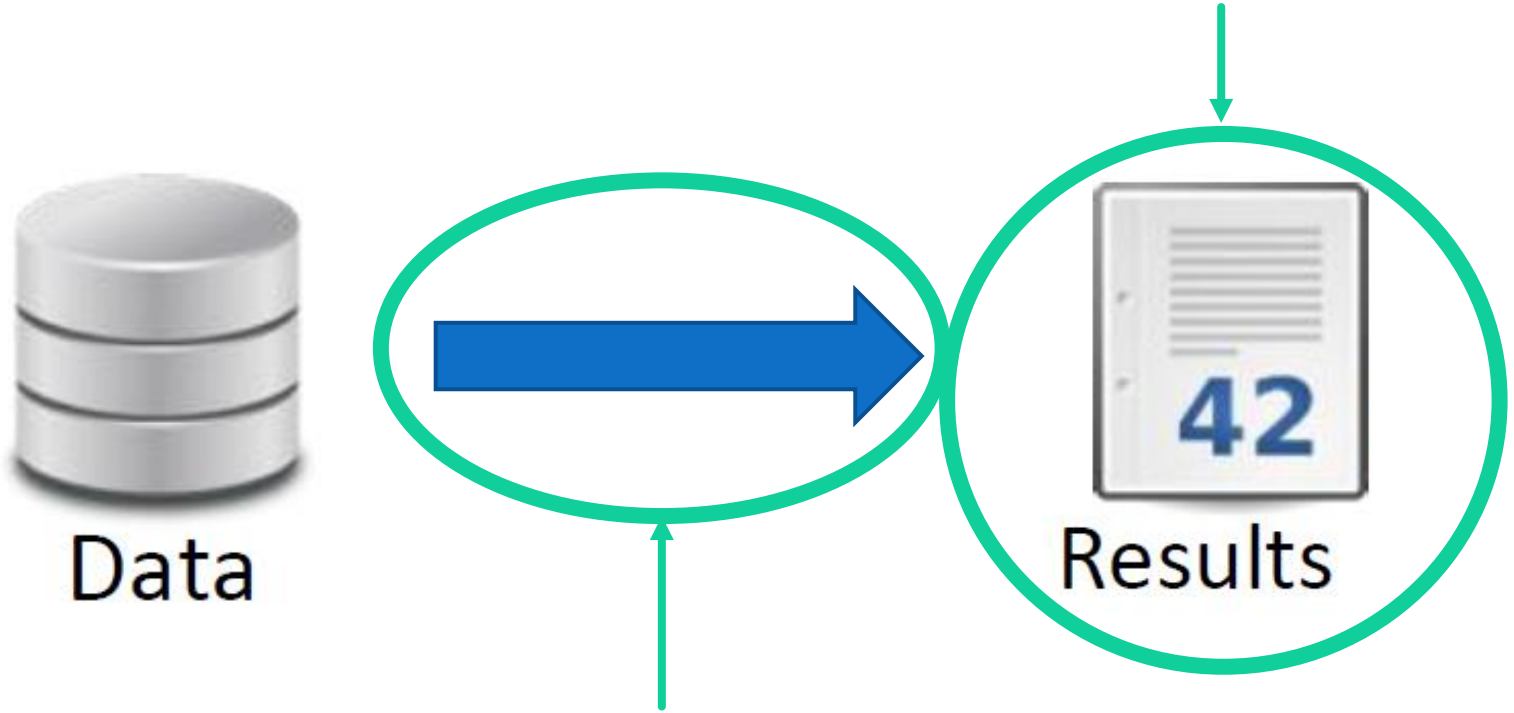
Problem: knowing $x + y$ reveals the other person's data

Still satisfies MPC's security definition

Differential Privacy (DP)

DP provides privacy-utility trade-off

DP protects results



MPC protects computation

MPC leaks zero knowledge beyond results (under crypto assumptions)

DP returns $x + y + noise$

x y

Scale of noise determines the privacy-utility trade-off

SQL under DP (SIGMOD 2021)

Dataset		TPC-H			Facebook				
Query		q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8
Query result		6,001,215	6,001,215	239,917	1,666,978,389	19,927	285,754	6,348,654	21,613
wPINQ output		8,680	9,770	385	54.2	131.4	44.1	23.8	183.6
Residual Sensitivity (RS)	Min Value	694	694	49	77,100,000	203	2,790	86,800	50
	Max Value	51,900	52,000	51,800	301,000,000	1,410	54,500	2,050,000	51,300
	Running Time(s)	27.6	53.6	48.6	2.96	1.68	4.71	18.1	20.1
Elastic Sensitivity (ES)	Min Value	1,740	2,140	175,000,000	25,500,000,000	219,000	110,000,000	55,000,000,000	561
	Max Value	4,950,000	1,870,000	263,000,000	25,500,000,000	219,000	110,000,000	55,000,000,000	2,440,000
	Running Time(s)	7.55	9.02	6.73	0.300	0.611	0.628	3.725	8.78
ES/RS	Min	2.51×	3.08×	5,069×	84.8×	156×	2,010×	26,900×	11.2×
	Max	273×	67×	3,580,000×	330×	1,080×	39,300×	634,000×	178×
	Avg	94.4×	27.8×	1,750,000×	286×	875×	27,300×	503,000×	80.6×

Reduction in noise scale under the same privacy guarantee,
Compared with [Johnson, Near, Song, 2018]

ML by SQL over private data
(Work in progress)

- Bank: R_1(person, gender, age, balance, …)
- Taobao: R_2(person, spending)

```
select LinearRegression(gender, age, balance, …, spending)
from R1, R2
where R1.person=R2.person
```

Thanks!