Privacy preserving computation: Use cases and problem statement

Ant Group, Hong Kong University of Science and Technology (Guangzhou)

2025.07

ANT GROUP | Alipay

The importance of data collaboration

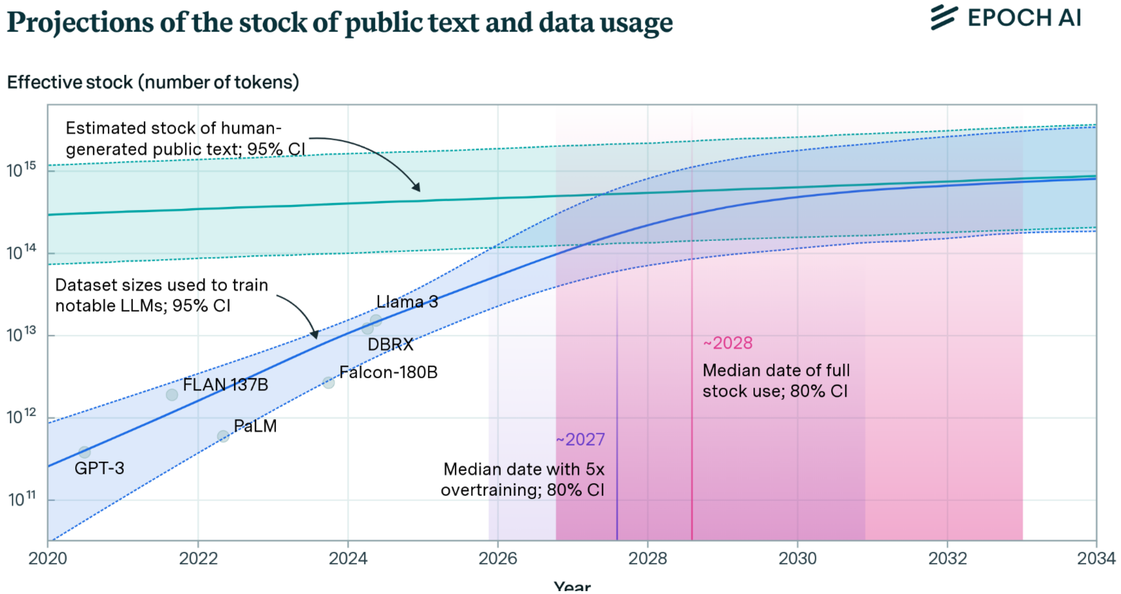
➢ Public-available data will soon be exhausted

- Companies will be at the same starting point if they use solely public data

➢ Proprietary data will be the next Blue Ocean market

- But they are often held by different organizations
- Securely Combining those proprietary data is the key to success

> **The Economist** ✓
> @TheEconomist
>
> AI firms will soon hit a “data wall” and exhaust the internet’s reserve of information. The stock of high-quality textual data on the internet will all have been used by 2028, one estimate finds.



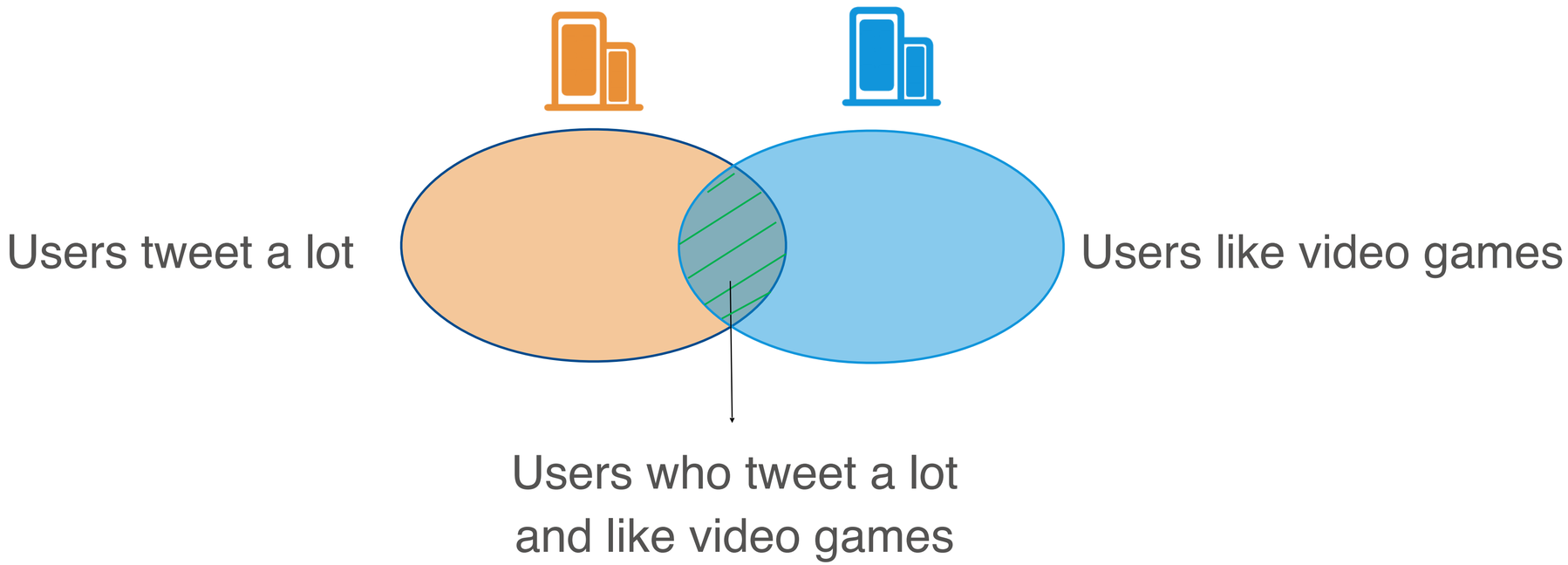
Use case 1: dataset intersection across companies

➢ Company A and Company B wants to find their mutual customers

- To establish a joint marketing campaign

➢ Reveal the (property of) mutual customers but **not any additional information**

- The Private Set Intersection (PSI) problem



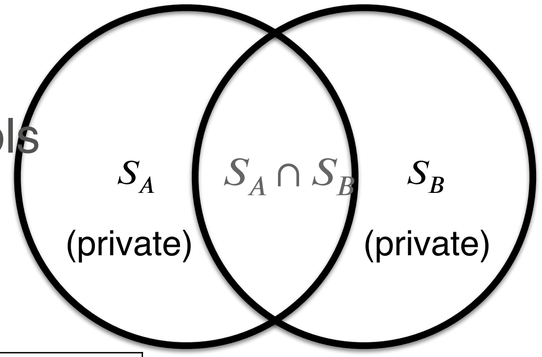
Use case 1: dataset intersection across companies

➢ Sending Hashed user ID is not enough

- User ID does not have enough entropy; Require specific cryptographic protocols

➢ Many related researches on PSI

- Based on Diffie-Hellman / Oblivious Transfer / VOLE ..



	Years since publication	Computation	Communication volume	Communication round
Diffie-Hellman based	~30	Low	Low	Low
Oblivious Transfer based	~10	Low	High	Moderate
VOLE-based	~3	Lowest	Low	Moderate

➢ ECDH-PSI have been widely deployed by Alibaba, Alipay, Bytedance, Bank of China …

- Fraud detection \ card activation \ issue coupons …

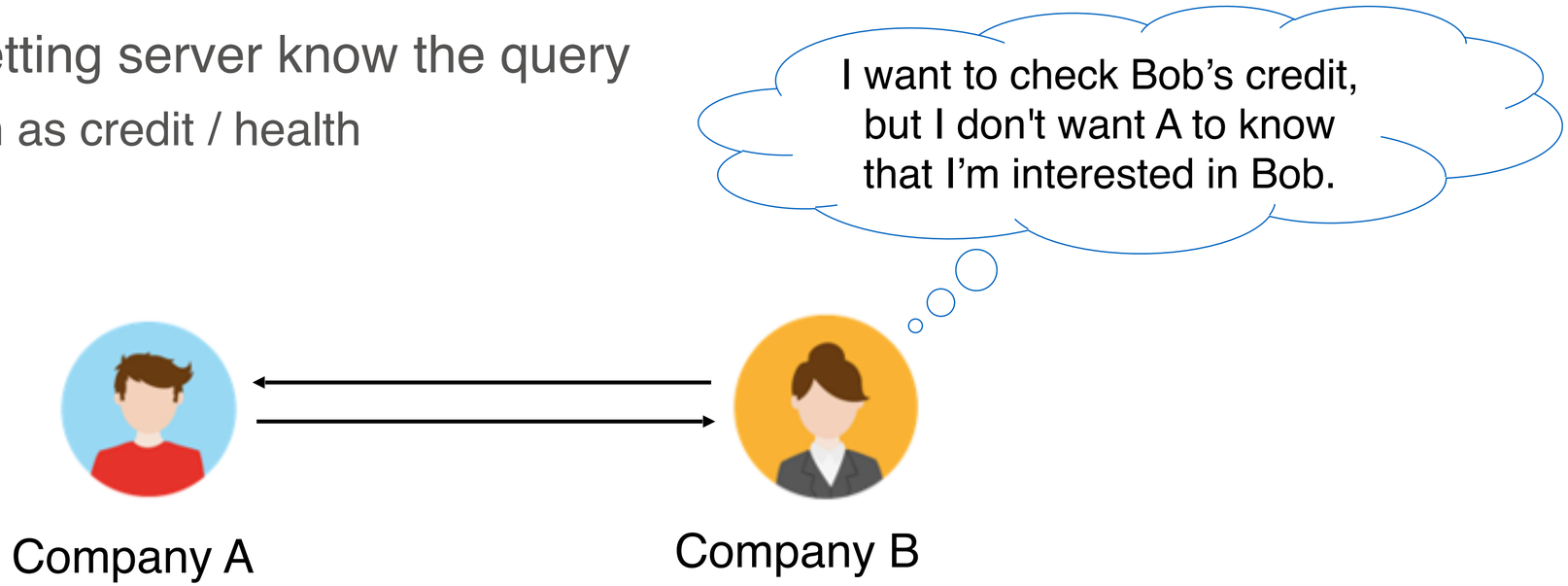
➢ An example use case

- “Chinese X” and “Chinese Youtube”, each holding ~1B PIDs, made an ECDH-PSI in ~half a day.
- Important business decisions were made according to the PSI output.

Use case 2: Private key-value query

➢ Making queries without letting server know the query

- For sensitive queries such as credit / health



Name	Credit
Alice	1111
Bob	2222
…	…

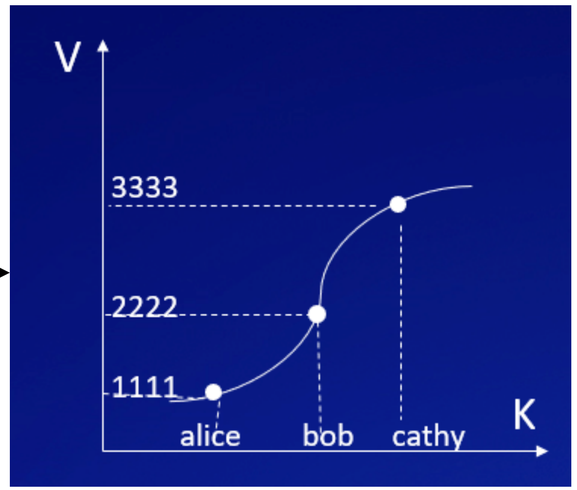
Use case 2: Private key-value query

- Traditionally, the database owner inevitably learns the row returned to the client
- But fully homomorphic encryption (FHE) could make private query possible
  - FHE allows addition/multiplication between encrypted values

Step 1: A interpolates her key-value database to obtain a polynomial: value = F(key).

Step 3: A homomorphically computes F on the encrypted key

Key	Value
Alice	1111
Bob	2222
Cathy	3333



Step 2: B sends encrypted key to A

Step 4: B decrypt and obtain the value

Enc(bob)

F(Enc(bob))

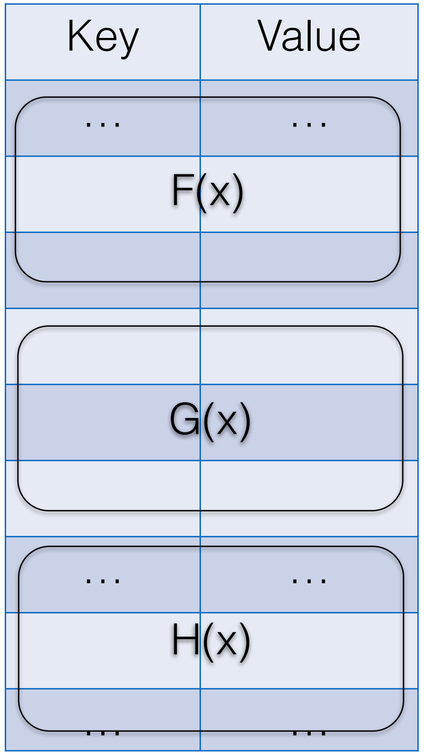
Use case 2: Private key-value query

➢ More optimizations

- Lower multiplication depth
- Batch queries
- Communication-computation trade-off

➢ An example use case in Alibaba Health: Medicine search system

- 100M + medicines, impractical for private query
- Split into 100 groups, each group uses different function
  // leaking which group to query
- Private query of 1M medicines in ~5 seconds.

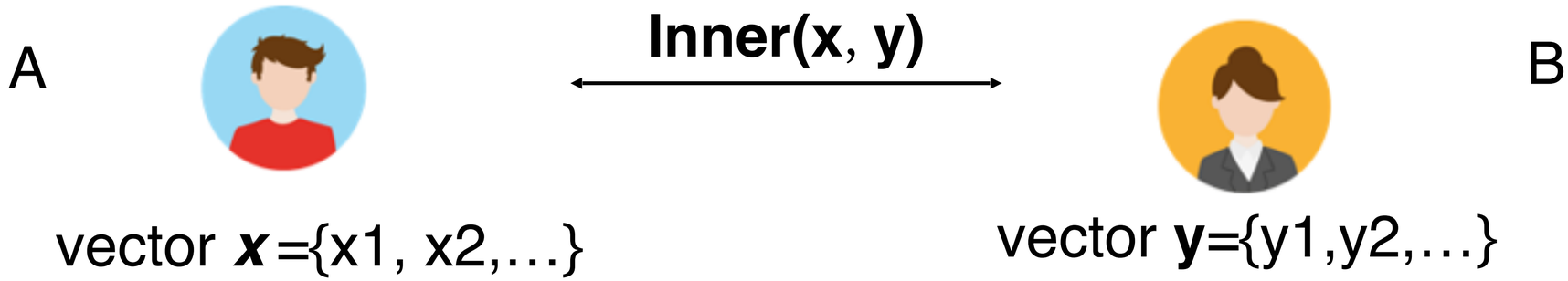


Use case 3: Private inner product

➢ Company A and Company B want to compute an inner product of their vectors

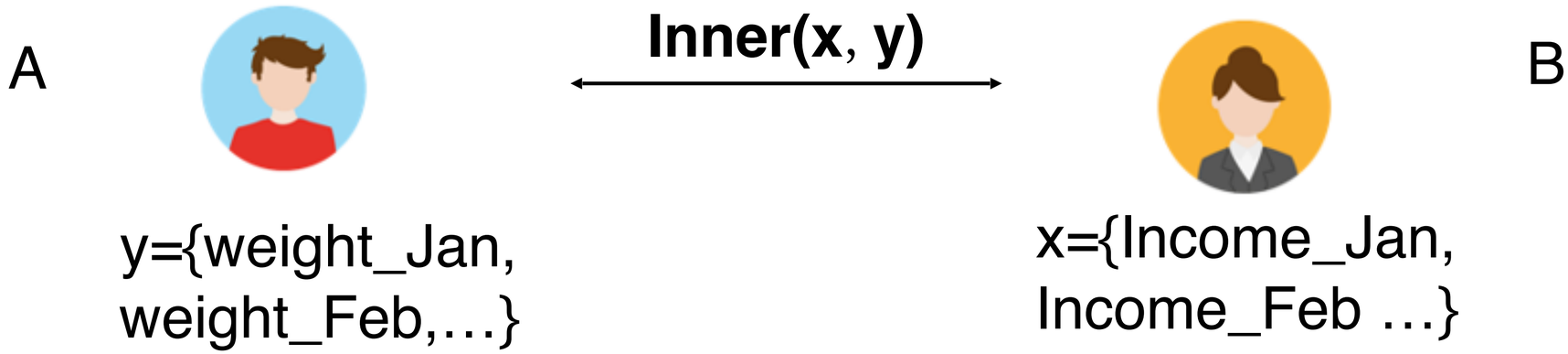
but **neither A nor B wants to share their vector directly**

- Customer Profile Similarity
- Genetic Relative Matching
- Suspect Fingerprint Matching



Use case 3: Private distance computation

- An example use case between “Big Bank A“ and “Small Bank B”
  - A has built a successful model : {Income each month-> credit}
  - B has no such model, but still wants to evaluate the credit of B’s customers
- Use additive homomorphic encryption Paillier
  - B send encrypt(x) to A; A compute encrypt($\Sigma x_i y_i$) and return to B; B decrypt
  - ~100miliseconds per evaluation



Use case 4: collaborative machine learning

➢ Company A and Company B want to jointly train a machine learning model, but **neither A nor B wants to directly share their raw data with the other**

- Fraud detection, credit assessment
- Personalized recommendation system

Company A  Train a model: Features > Label  Company B

PID	Feature1	Feature2	Feature3
Alice			
…	…		…

PID	Feature4	Label
Alice		
…	…	

Use case 4: collaborative machine learning

- An example use case between “Chinese Amazon” B and “Chinese Verizon” A
- B found some (potential) fraud users, want to make sure with help of A
  - Step 1: Each company clean its own dataset
  - Step 2: Do a PSI to align the dataset
  - Step 3: Run a secure XGBoost training using MPC.
  - ~10M users, 10 features finish in ~ one day [1]

A  Is Cathy a fraud?  B

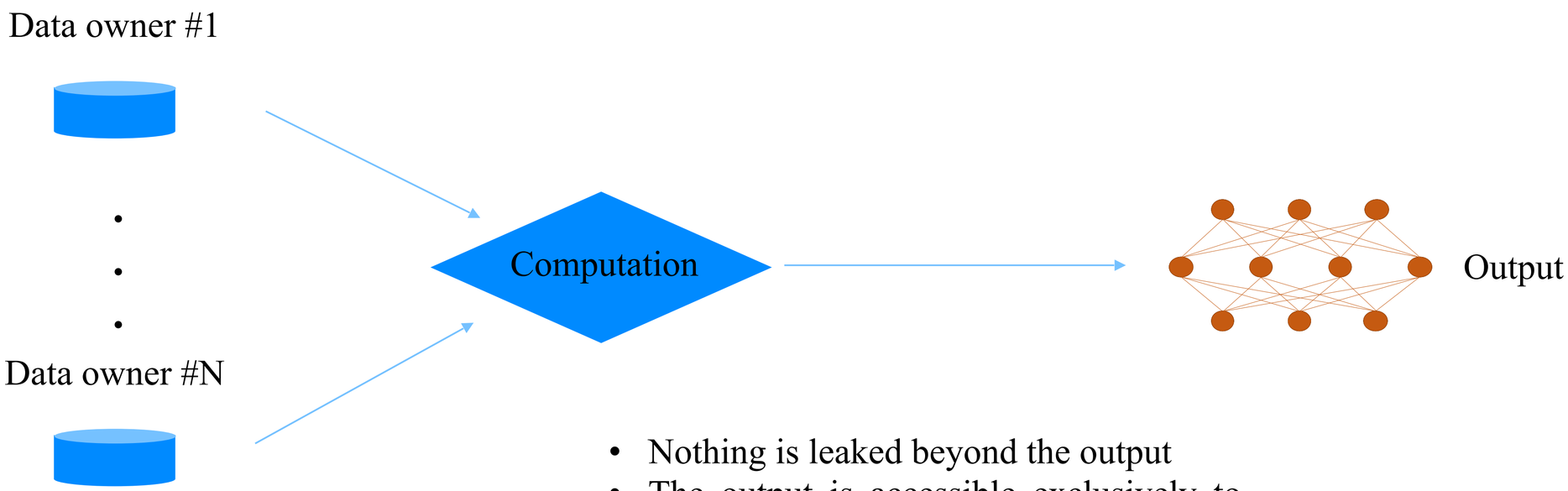
PID	Calls per day	Avg. Call length	Avg. Plan fee
Alice			
Bob	…		…
Cathy			

PID	Goods Bought	IsFraud
Alice		Yes
Bob	…	No
Cathy		?

[1] Squirrel: A Scalable Secure Two-Party Computation Framework for Training Gradient Boosting Decision Tree, USENIX Security 2023

Problem statement

➢ A few data owners collaboratively run a function on the joint input



- Nothing is leaked beyond the output
- The output is accessible exclusively to **authorized party**

Problem statement

➢Security

- **Most of the time, semi-honest security is enough.** Parties are not randomly picked, but are companies that already have incentive to collaborate. (Rationally) They won't do malicious behaviors because of reputation.

➢Functionality

- **Algorithms support.** Solutions should support multiple tasks, such as set intersection, joint average/sum/max, joint modeling…, to meet needs in different business scenarios
- **User-friendly operations.** Solutions should offer easy-to-use APIs to lower the usage barrier and enhance user experience

➢Performance

- **Computation overhead.** Solutions should make best use of available computational power
- **Communication overhead.** Solutions should minimize the amount and round of data transmitted across networks

Proposed next steps

- IETF/IRTF is dedicated to furthering internet standards that foster secure and efficient data collaboration. **We suggest commencing research and standardization efforts on Privacy-Preserving computation (PPC) technologies for cross-companies data collaboration.**
- Privacy preserving technologies enable multiple parties to jointly perform data computation while safeguarding security. They allow for collaborative data processing without unauthorized disclosure to computation parties or external entities:
  - **Private set intersection (PSI)** **draft-wang-ppm-ecdh-psi-01**
  - Homomorphic Encryption (HE)
  - Multi-Party Computation (MPC)
  - Federated Learning (FL)
  - Trusted Execution Environment (TEE)

ANT GROUP | Alipay

Ongoing researches

➢ Hot research topics concerned by security/cryptographic academy and industry

➢ Private Set Intersection (PSI)

- CCS 2022“Blazing Fast PSI from Improved OKVS and Subfield VOLE”
- USENIX Security 2023 “Distance-Aware Private Set Intersection”

➢ Privacy-Preserving Query (also known as Private Information Retrieval (PIR), or labeled-PSI)

- CCS 2018“Labeled PSI from Fully Homomorphic Encryption with Malicious Security”
- CCS 2024“Faster FHE-Based Single-Server Private Information Retrieval”
- USENIX Security 2024 “Single Pass Client Preprocessing Private Information Retrieval”

Ongoing researches

➢ Privacy-Preserving Machine Learning (PPML)

- USENIX Security 2022 “Cheetah: Lean and Fast Secure Two-Party Deep Neural Network Inference”
- USENIX Security 2023 “Squirrel: A Scalable Secure Two-Party Computation Framework for Training Gradient Boosting Decision Tree”
- NDSS 2025 “Bumblebee: Secure Two-Party Inference Framework for Large Transformers”
- USENIX Security 2023 “Efficient 3PC for Binary Circuits with Applications to Malicious-Secure DNN Inference”
- CCS 2024 “Coral: Maliciously Secure Computation Framework for Packed and Mixed Circuits”
- S&P 2024 “Don’t Eject the Impostor: Fast Three-Party Computation With a Known Cheater”
- CCS 2022 “pMPL: A Robust Multi-Party Learning Framework with a Privileged Party”


THE END
THANK YOU!
ANT GROUP
Alipay