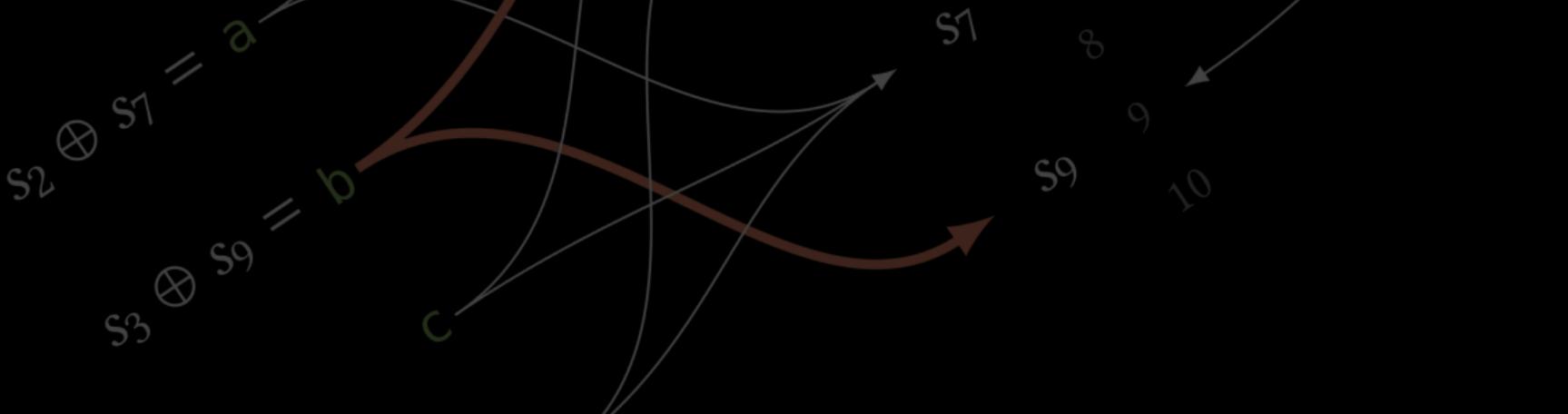


A Brief Overview of **Private Set Intersection**

Mike Rosulek, Oregon State University

NIST STPPA, April 19, 2021



what is private set intersection (PSI)?

Alice

p x o

n r e

s u m

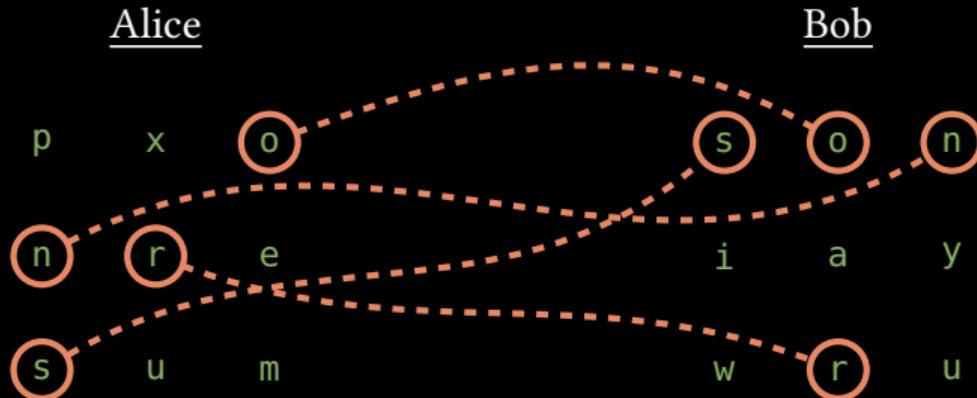
Bob

s o n

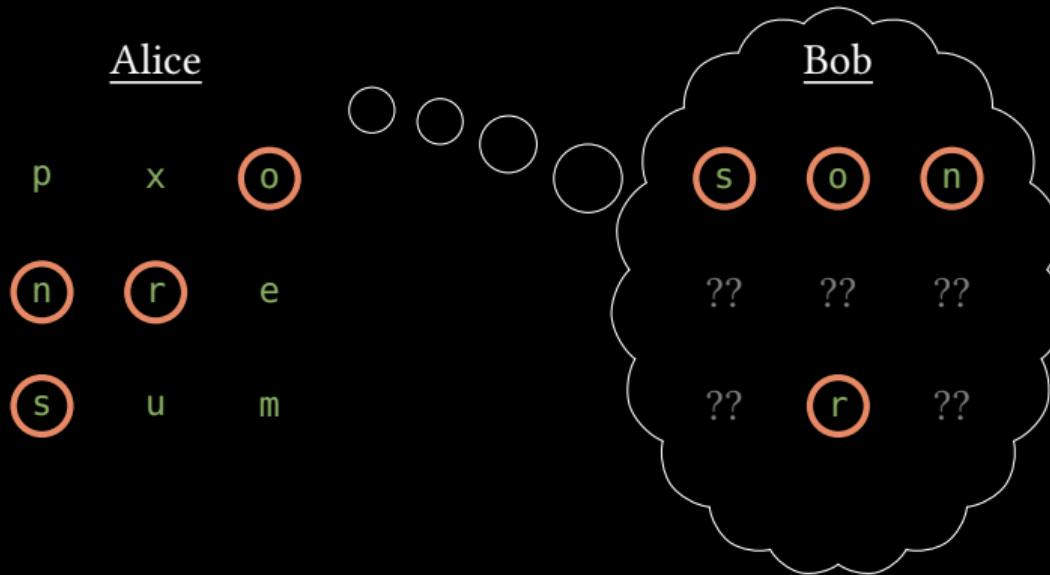
i a y

w r u

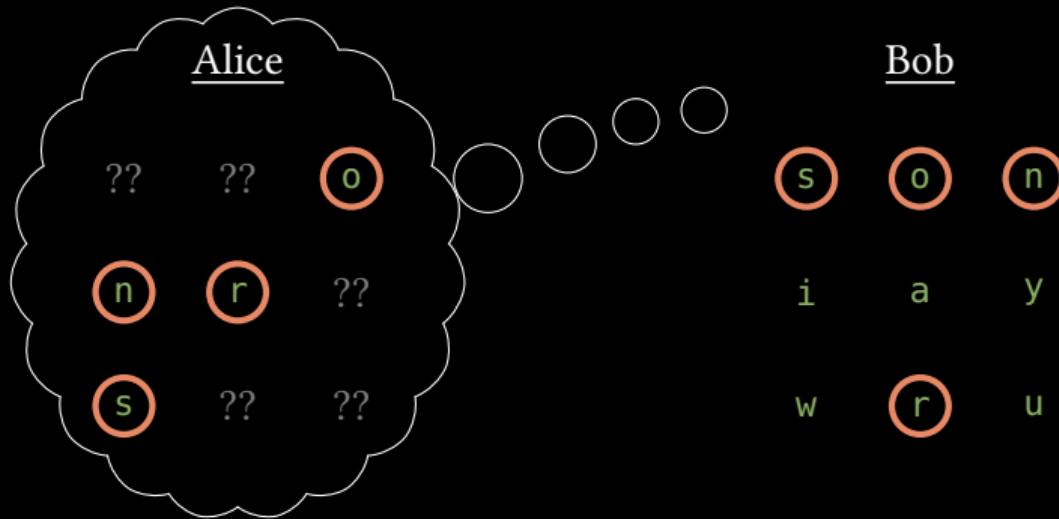
what is private set intersection (PSI)?



what is private set intersection (PSI)?



what is private set intersection (PSI)?



why use PSI?

The Difficulty Of Private Contact Discovery

moxie0 on 03 Jan 2014

Building a social network is not easy. Social networks have value proportional to their size, so participants aren't motivated to join new social networks which aren't already large. It's a paradox where if people haven't already joined, people aren't motivated to join.

{my phone contacts} \cap {users of your service}

why use PSI?

The Difficulty Of Private Contact Discovery

moxie0 on 03 Jan 2014

network is not easy. Social networks have value
their size, so participants aren't motivated to join new
which aren't already large. It's a paradox where if
ready joined, people aren't motivated to join.

Google's Password Checkup feature will be built into Chrome

The tool warns users if their passwords are known to be compromised.

G. Torbet
@georjinatorbet
October 2nd, 2019

{my passwords} \cap {passwords found in breaches}

why use PSI?

The collage consists of three screenshots:

- Top Left:** A screenshot of a Signal messaging interface showing a group availability poll. It includes a photo of a ship and the text "The Difficulty Of Private C".
- Top Right:** A screenshot of a calendar-based availability poll from "moxie0" dated Jan 3, 2014. It shows a grid of participants (2, Janet, Jose) across dates (May 2-4) and times (8:00 PM - 9:00 PM, 10:00 PM - 11:00 PM). Green checkmarks indicate availability.
- Bottom Left:** A screenshot from Engadget titled "Google's Password Checkup feature will be built into Chrome". It features a photo of a person holding a smartphone and a laptop, with the text "The tool warns users if their passwords are known to be compromised." and author details: "G. Torbet @georjinatorbet October 2nd, 2019".

{my availability} \cap {your availability}

why use PSI?

The screenshot shows a Signal app interface. At the top is a blue header with the Signal logo. Below it is a white message area containing a circular profile picture of a ship at sea. To the right is a calendar grid titled "The Difficulty Of Private C". The grid shows a week from May 2nd to May 4th. It includes columns for "May 2 THU" (8:00 PM - 10:00 PM), "May 3 FRI" (7:00 PM - 9:00 PM), "May 3 FRI" (9:00 PM - 11:00 PM), "May 4 SAT" (7:00 PM - 9:00 PM), and "May 4 SAT" (9:00 PM - 11:00 PM). There are five rows: "2 participants" (blue header), "moxie0" (blue row), "janet" (green row), "jose" (green row), and a summary row with a green checkmark and a "Send" button.

engadget

Google's Password Checkup feature will be built into Chrome

The tool warns users if their passwords are known to be compromised.

G. Torbet
@georinatrbet
October 2nd, 2019

WIRED BUSINESS CULTURE DEAR IDEAS SCIENCE SECURITY TRANSPORTATION SIGN IN SUBSCRIBE

LILY HAY KEMPER SECURITY 08.19.2019 09:00 AM

Google Turns to Retro Cryptography to Keep Data Sets Private

Google's Private Join and Compute will let companies compare notes without divulging sensitive information.

{people who saw ad} \cap {customers who made purchases}

why use PSI?

The New York Times

Who Is Registered to Vote in Two States? Some in Trump's Inner Circle

Log In

moxie0 on 03 Jan 2014

Private C

| | May 2 THU | May 3 FRI | May 3 FRI | May 4 SAT | May 4 SAT |
|---------------------|---------------------|---------------------|--------------------|---------------------|--------------------|
| 8:00 PM 10:00 PM | 9:00 PM 11:00 PM | 9:00 PM 11:00 PM | 7:00 PM 9:00 PM | 9:00 PM 11:00 PM | 7:00 PM 9:00 PM |
| 2 participants | ✓0 | ✓1 | ✓1 | ✓2 | ✓2 |
| moxie0 | ✓ | ✓ | ✓ | ✓ | ✓ |
| janet | | | ✓ | ✓ | ✓ |
| jose | | ✓ | | ✓ | ✓ |

Send

engadget

Google's Password Checkup feature will be built into Chrome

The tool warns users if their passwords are known to be compromised.

G. Torbet
@georinotorbet
October 2nd, 2019

WIR ED

network is not easy. Social networks have value
their size, so
which aren't
ready joined

LILY HAY KENNEDY SECURITY 08.19.2019 09:00 AM

Google Turns to Retro Cryptography to Keep Data Sets Private

Google's Private Join and Compute will let companies compare notes without divulging sensitive information.

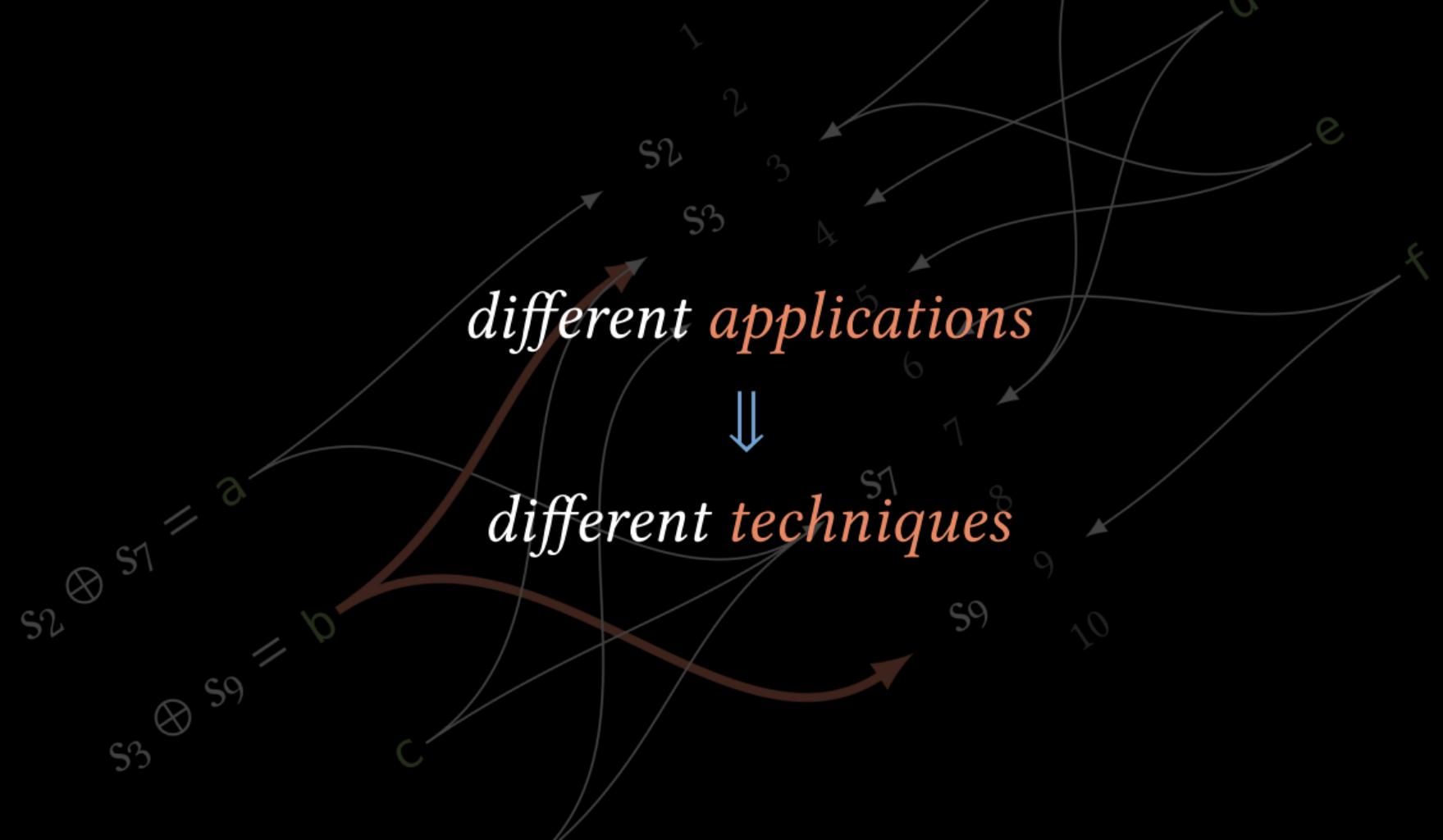
{voters registered in OR} \cap {voters registered in NY}

why use PSI?

The collage consists of four screenshots:

- Top Left:** A screenshot of a mobile application interface. It features a large red semi-circle at the top with the text "ELECTRONIC REGISTRATION" and a blue semi-circle at the bottom with "INFORMATION CENTER". In the center, there is a logo with two stars and the word "ERIC". To the left, the text "Who Is In My States?" is visible. Below the logo, there are social media sharing icons for Facebook, Twitter, and LinkedIn.
- Top Right:** A screenshot of a calendar application showing a weekly schedule from May 2nd to May 4th. The schedule includes time slots from 8:00 PM to 9:00 PM. A grid shows participants' availability: "2 participants" (blue row) has checkmarks in all slots; "janet" (green row) has checkmarks in the 9:00 PM slots; and "jose" (green row) has checkmarks in the 8:00 PM and 9:00 PM slots. A green button labeled "Send" is at the bottom right.
- Bottom Left:** A screenshot from Engadget's website. The headline reads "Google's Password Checkup feature will be built into Chrome". Below the headline, it says "The tool warns users if their passwords are known to be compromised." There is a photo of a person holding a smartphone next to a coffee cup.
- Bottom Right:** A screenshot from WIRED magazine's website. The headline reads "Google Turns to Retro Cryptography to Keep Data Sets Private". Below the headline, it says "Google's Private Join and Compute will let companies compare notes without divulging sensitive information." The WIRED logo and navigation menu are visible at the top.

{voters registered in OR} \cap {voters registered in NY}





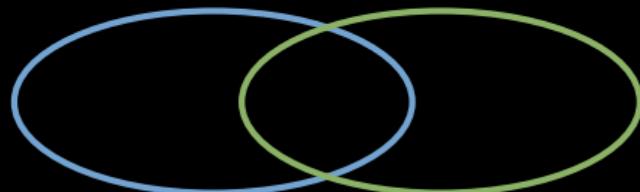
PSI on **small sets** (hundreds)

- ▶ private availability poll
- ▶ key agreement techniques



PSI on **small sets** (hundreds)

- ▶ private availability poll
- ▶ key agreement techniques



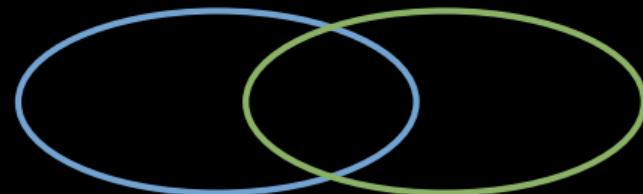
PSI on **large sets** (millions)

- ▶ double-registered voters
- ▶ OT extension; combinatorial tricks



PSI on **small sets** (hundreds)

- ▶ private availability poll
- ▶ key agreement techniques



PSI on **large sets** (millions)

- ▶ double-registered voters
- ▶ OT extension; combinatorial tricks



PSI on **asymmetric sets** (100 : billion)

- ▶ contact discovery; password checkup
- ▶ offline phase; leakage



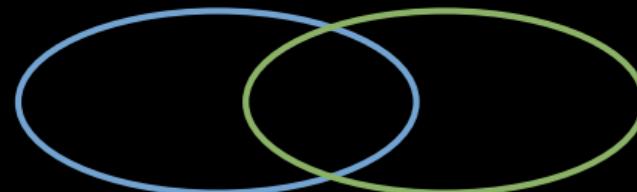
PSI on **small sets** (hundreds)

- ▶ private availability poll
- ▶ key agreement techniques



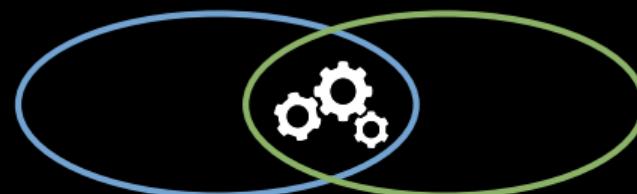
PSI on **asymmetric sets** (100 : billion)

- ▶ contact discovery; password checkup
- ▶ offline phase; leakage



PSI on **large sets** (millions)

- ▶ double-registered voters
- ▶ OT extension; combinatorial tricks



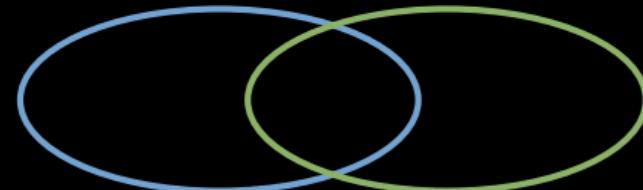
computing on the intersection

- ▶ sales statistics about intersection
- ▶ generic MPC



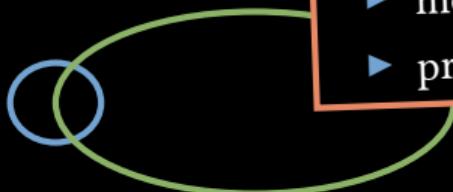
PSI on **small sets** (hundreds)

- ▶ private availability poll
- ▶ key agreement techni



PSI on **large sets** (millions)

- ▶ double registered voters
- ▶; combinatorial tricks



PSI on **asymmetric sets** (100 : billion)

- ▶ contact discovery; password checkup
- ▶ offline phase; leakage

Not to mention:

- ▶ approximate/fuzzy matching
- ▶ more than 2 parties/sets
- ▶ private set *union*



computing on the intersection

- ▶ sales statistics about intersection
- ▶ generic MPC

PSI on small sets

key agreement techniques

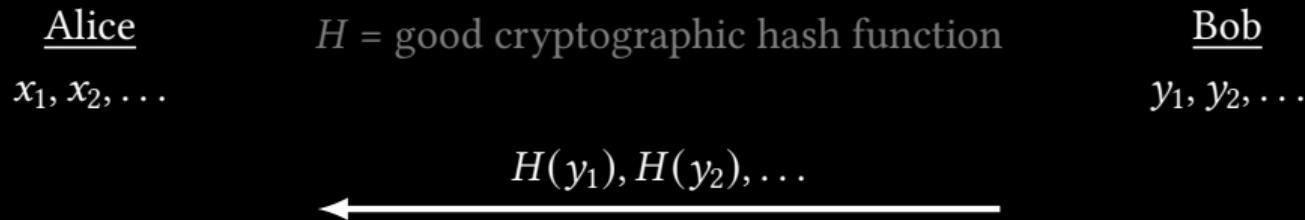
$$S_2 \otimes S_1 = a$$

$$S_3 \otimes S_9 = b$$

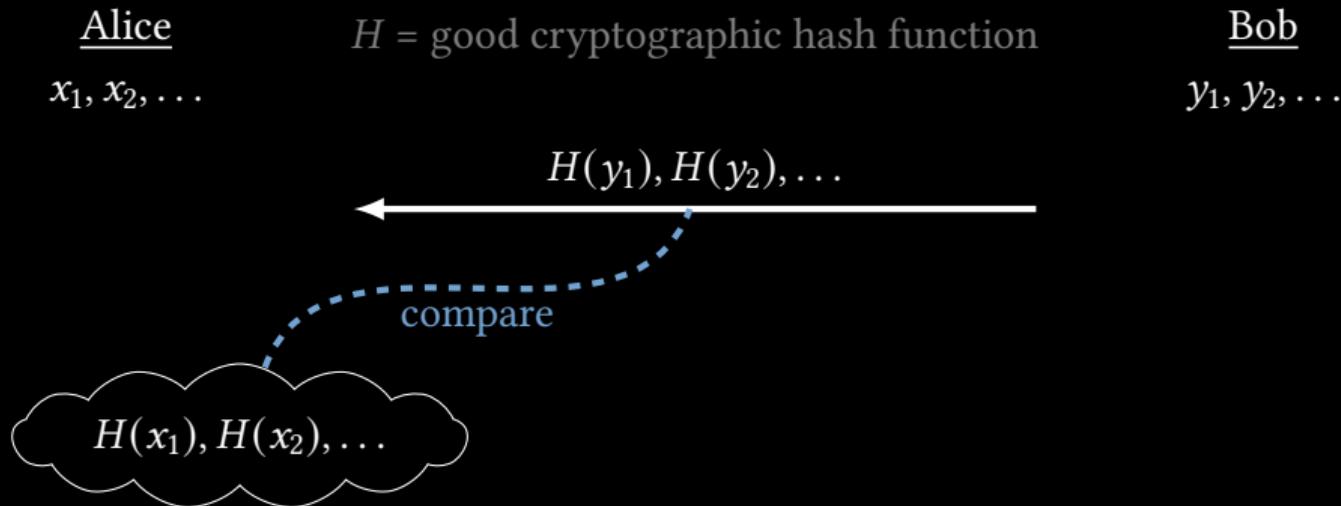
c



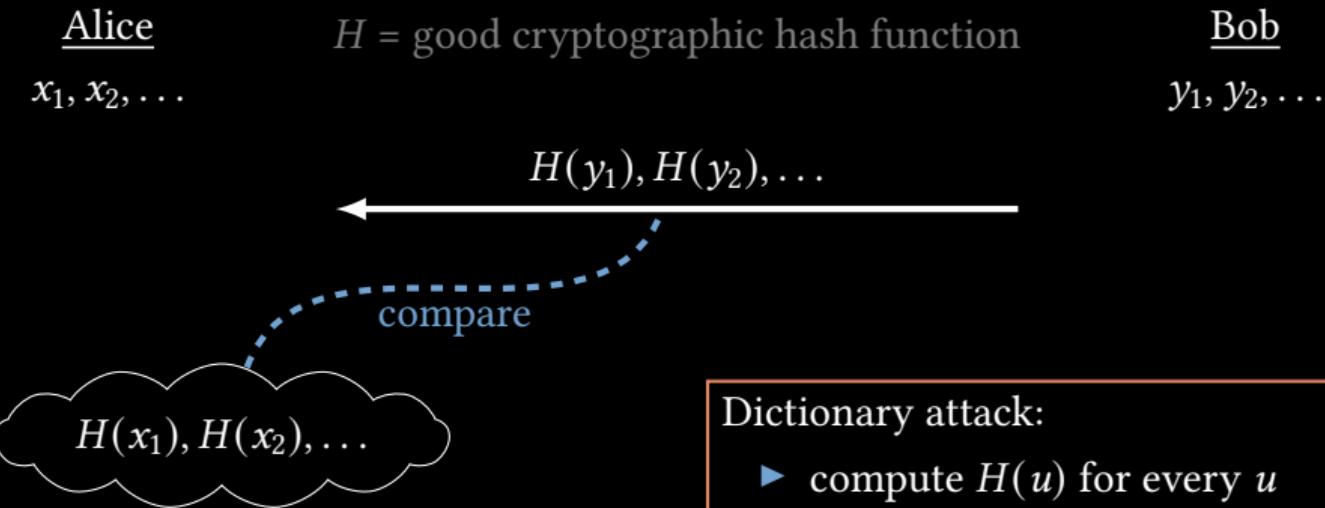
*a **bad** mental model for PSI*



*a **bad** mental model for PSI*



*a **bad** mental model for PSI*



*a **bad** mental model for PSI*

Alice

x_1, x_2, \dots

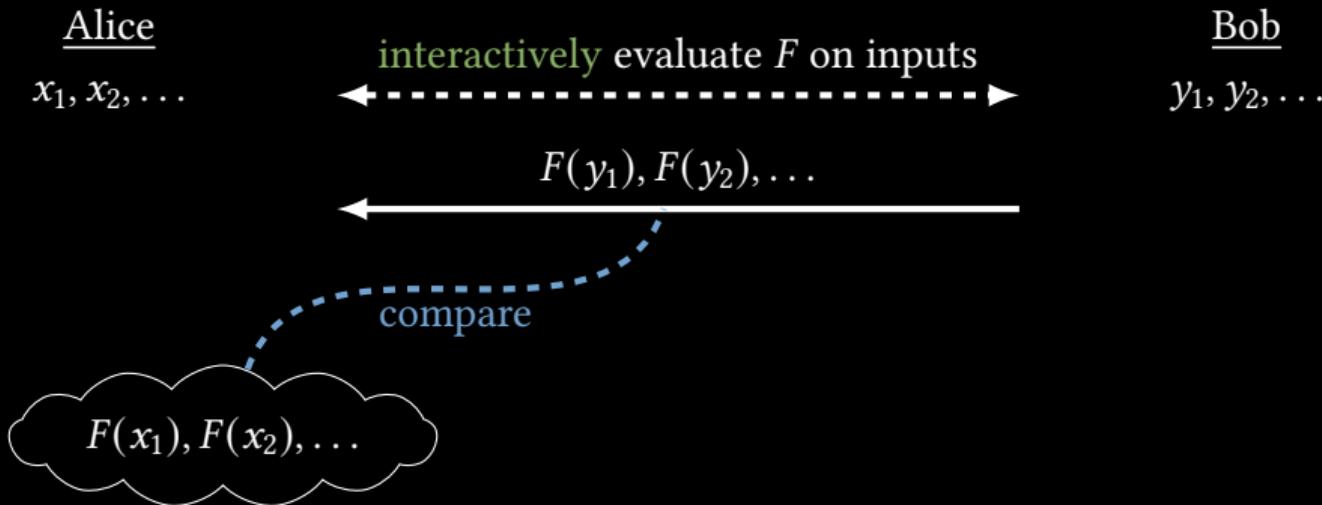
$H = \text{good cryptographic hash function}$



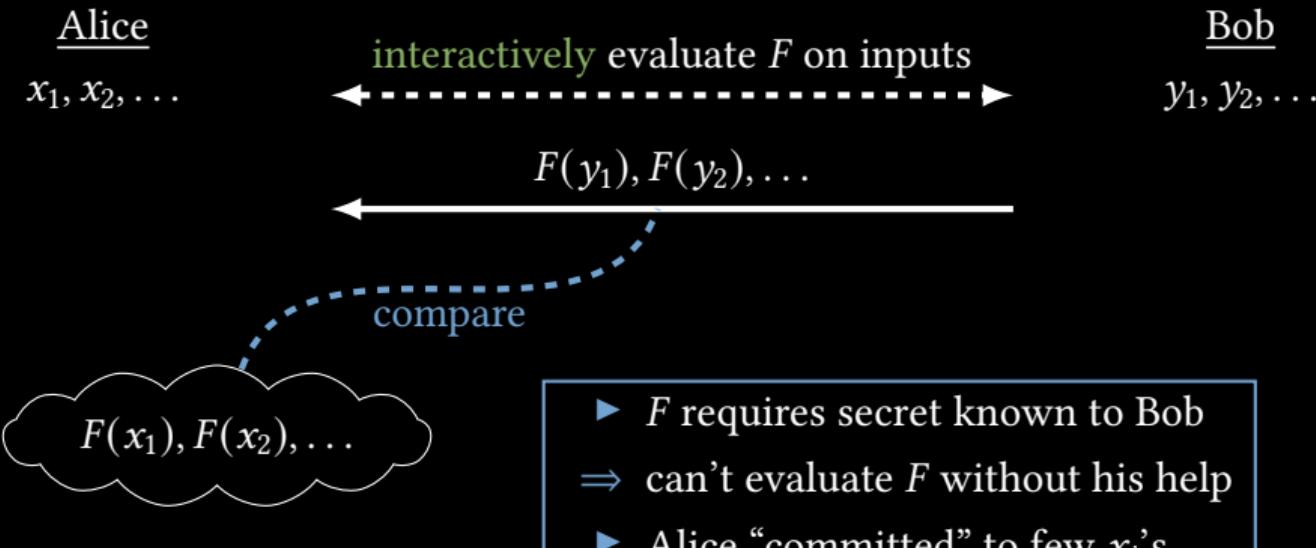
Bob

y_1, y_2, \dots

*a **better** mental model for PSI*



*a **better** mental model for PSI*



Alice

x

Does $x = y?$

Bob

y

[Shamir80, Meadows86, Jablon96]

Alice

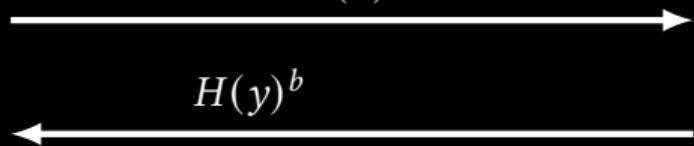
x

$H = \text{random oracle}$

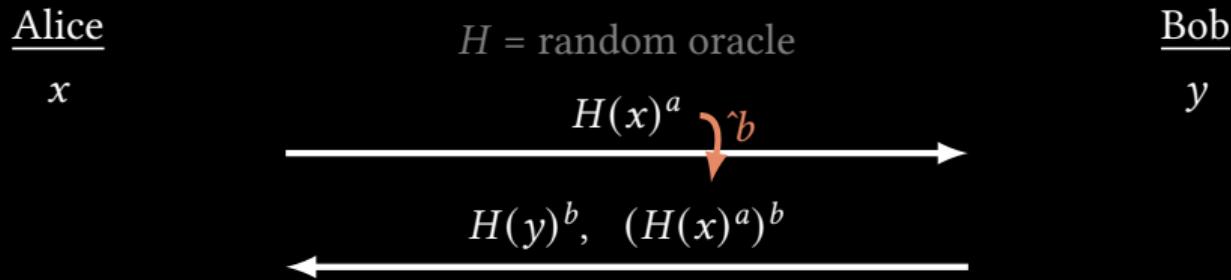
$H(x)^a$

Bob

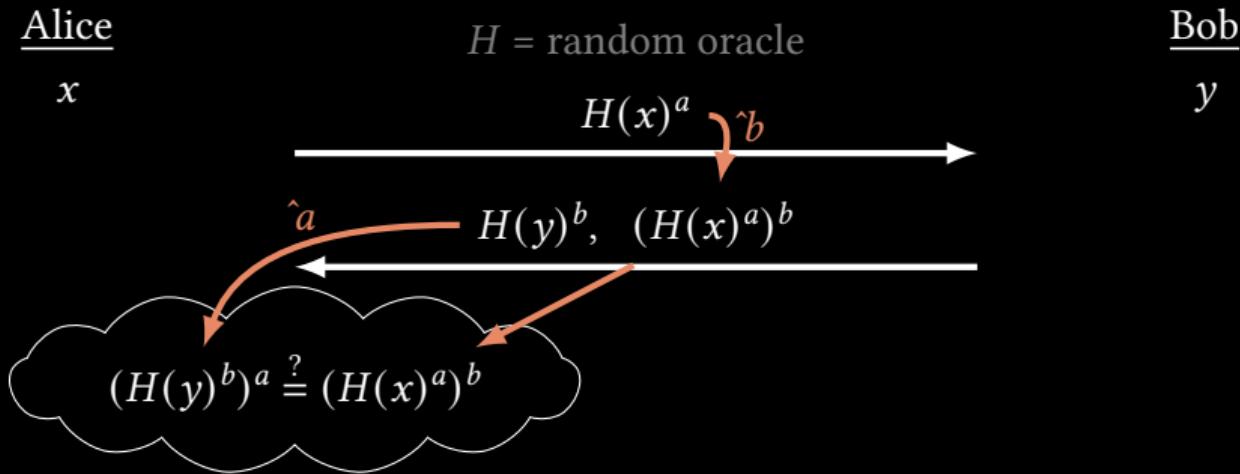
y



[Shamir80, Meadows86, Jablon96]



[Shamir80, Meadows86, Jablon96]



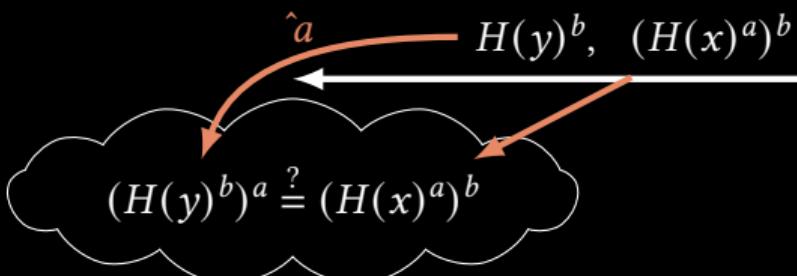
[Shamir80, Meadows86, Jablon96]

Alice

x

$H = \text{random oracle}$

$$H(x)^a \xrightarrow{\quad} \hat{y}^b$$

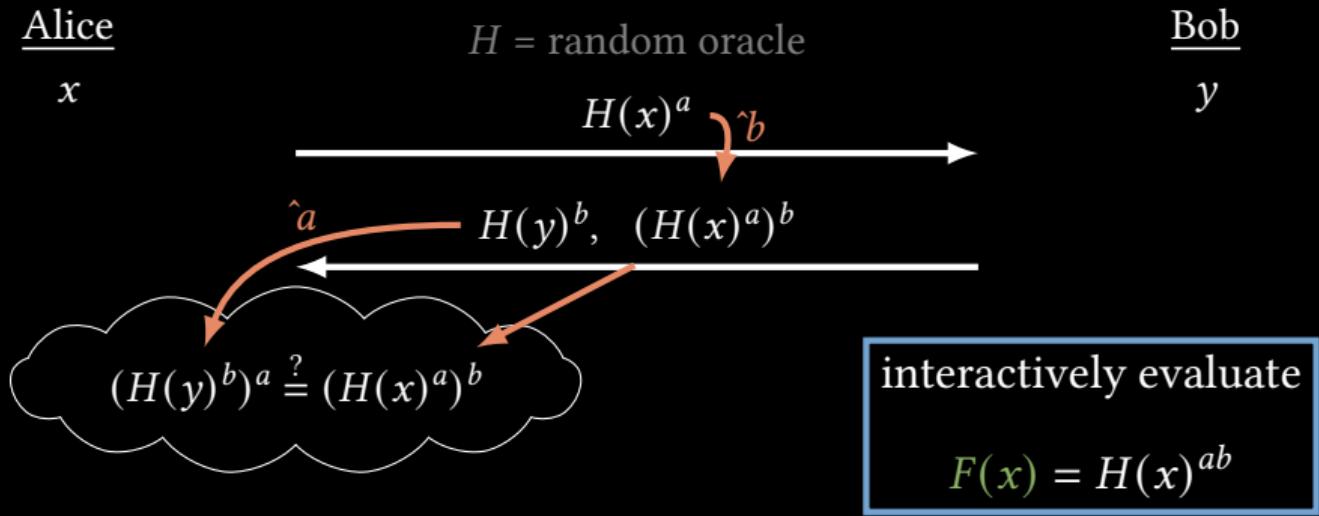


Bob

y

interactively evaluate
 $F(x) = H(x)^{ab}$

[Shamir80, Meadows86, Jablon96]



$x \neq y \xrightarrow{\text{RO}} H(y)$ independent of everything else $\xrightarrow{\text{DDH}}$ $H(y)^b \approx \$$

[Shamir80, Meadows86, Jablon96]

Alice

x_1, x_2, \dots

Bob

y_1, y_2, \dots

What is $X \cap Y$?

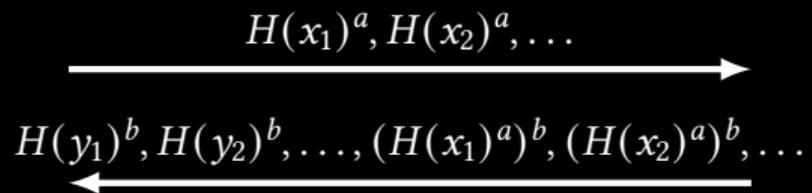
[HubermanFranklinHogg99]

Alice

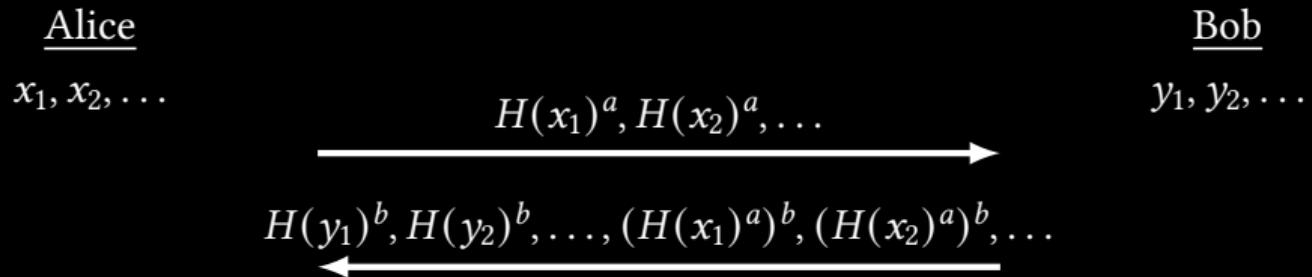
x_1, x_2, \dots

Bob

y_1, y_2, \dots



[HubermanFranklinHogg99]



[HubermanFranklinHogg99]

- ▶ Malicious security via ZK [DeCristofaroKimTsudik10,JareckiLiu09]
- ▶ Authenticated items [DeCristofaroKimTsudik10]
- ▶ From generic key agreement [RosulekTrieu21]

overview: PSI on small sets

for 256 items:
0.1 seconds; 10 KB

with malicious security!

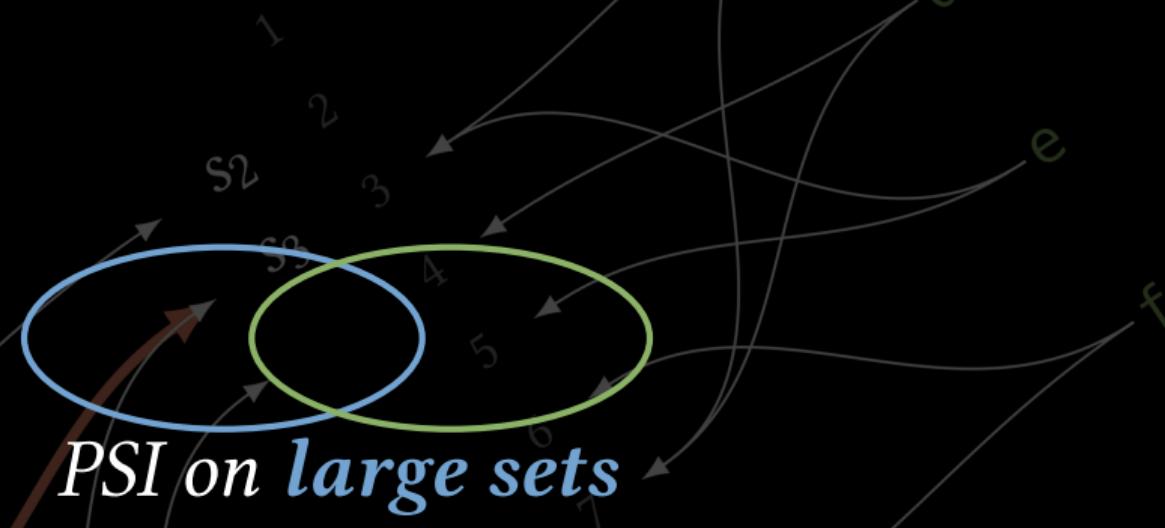
PSI on large sets

OT & hashing techniques; scaling to 1M items

$$S_2 \otimes S_1 = a$$

$$S_3 \otimes S_9 = b$$

c



scaling to 1 million items?

$$H(x_1)^a, H(x_2)^a, \dots, H(x_{1000000})^a$$


scaling to 1 million items?

$$\xrightarrow{\underbrace{H(x_1)^a, H(x_2)^a, \dots, H(x_{1000000})^a}_{> 4 \text{ minutes!}}}$$

batch oblivious PRF (OPRF)

Alice

1

2

3

4

5

6

7

8

9

⋮

Bob

batch oblivious PRF (OPRF)

Alice

$x_1 \quad 1$

$x_2 \quad 2$

$x_3 \quad 3$

$x_4 \quad 4$

$x_5 \quad 5$

$x_6 \quad 6$

$x_7 \quad 7$

$x_8 \quad 8$

$x_9 \quad 9$

Bob

\vdots

batch oblivious PRF (OPRF)

Alice

$$\mathbf{F}_1(x_1) \quad 1 \quad \mathbf{F}_1(\cdot)$$

$$\mathbf{F}_2(x_2) \quad 2 \quad \mathbf{F}_2(\cdot)$$

$$\mathbf{F}_3(x_3) \quad 3 \quad \mathbf{F}_3(\cdot)$$

$$\mathbf{F}_4(x_4) \quad 4 \quad \mathbf{F}_4(\cdot)$$

$$\mathbf{F}_5(x_5) \quad 5 \quad \mathbf{F}_5(\cdot)$$

$$\mathbf{F}_6(x_6) \quad 6 \quad \mathbf{F}_6(\cdot)$$

$$\mathbf{F}_7(x_7) \quad 7 \quad \mathbf{F}_7(\cdot)$$

$$\mathbf{F}_8(x_8) \quad 8 \quad \mathbf{F}_8(\cdot)$$

$$\mathbf{F}_9(x_9) \quad 9 \quad \mathbf{F}_9(\cdot)$$

⋮

Bob

batch oblivious PRF (OPRF)

Alice

$$\mathbf{F}_1(x_1) \quad 1 \quad \mathbf{F}_1(\cdot)$$

$$\mathbf{F}_2(x_2) \quad 2 \quad \mathbf{F}_2(\cdot)$$

$$\mathbf{F}_3(x_3) \quad 3 \quad \mathbf{F}_3(\cdot)$$

$$\mathbf{F}_4(x_4) \quad 4 \quad \mathbf{F}_4(\cdot)$$

$$\mathbf{F}_5(x_5) \quad 5 \quad \mathbf{F}_5(\cdot) \quad \text{learns nothing about } x_i \text{'s}$$

$$\mathbf{F}_6(x_6) \quad 6 \quad \mathbf{F}_6(\cdot)$$

$$\mathbf{F}_7(x_7) \quad 7 \quad \mathbf{F}_7(\cdot)$$

$$\mathbf{F}_8(x_8) \quad 8 \quad \mathbf{F}_8(\cdot)$$

$$\mathbf{F}_9(x_9) \quad 9 \quad \mathbf{F}_9(\cdot)$$

⋮

Bob

batch oblivious PRF (OPRF)

Alice

| | | | | | |
|---|--|---------------------|---|-----------------------|-------------------------------|
| | | $\mathbf{F}_1(x_1)$ | 1 | $\mathbf{F}_1(\cdot)$ | |
| | | $\mathbf{F}_2(x_2)$ | 2 | $\mathbf{F}_2(\cdot)$ | |
| | | $\mathbf{F}_3(x_3)$ | 3 | $\mathbf{F}_3(\cdot)$ | |
| | | $\mathbf{F}_4(x_4)$ | 4 | $\mathbf{F}_4(\cdot)$ | |
| all other $\mathbf{F}_i(x^*)$ look random | | $\mathbf{F}_5(x_5)$ | 5 | $\mathbf{F}_5(\cdot)$ | learns nothing about x_i 's |
| | | $\mathbf{F}_6(x_6)$ | 6 | $\mathbf{F}_6(\cdot)$ | |
| | | $\mathbf{F}_7(x_7)$ | 7 | $\mathbf{F}_7(\cdot)$ | |
| | | $\mathbf{F}_8(x_8)$ | 8 | $\mathbf{F}_8(\cdot)$ | |
| | | $\mathbf{F}_9(x_9)$ | 9 | $\mathbf{F}_9(\cdot)$ | |
| | | | | | \vdots |

Bob

batch oblivious PRF (OPRF)

Alice

| | | | | | |
|---|--|---------------------|---|-----------------------|-------------------------------|
| | | $\mathbf{F}_1(x_1)$ | 1 | $\mathbf{F}_1(\cdot)$ | |
| | | $\mathbf{F}_2(x_2)$ | 2 | $\mathbf{F}_2(\cdot)$ | |
| | | $\mathbf{F}_3(x_3)$ | 3 | $\mathbf{F}_3(\cdot)$ | |
| | | $\mathbf{F}_4(x_4)$ | 4 | $\mathbf{F}_4(\cdot)$ | |
| all other $\mathbf{F}_i(x^*)$ look random | | $\mathbf{F}_5(x_5)$ | 5 | $\mathbf{F}_5(\cdot)$ | learns nothing about x_i 's |
| | | $\mathbf{F}_6(x_6)$ | 6 | $\mathbf{F}_6(\cdot)$ | |
| | | $\mathbf{F}_7(x_7)$ | 7 | $\mathbf{F}_7(\cdot)$ | |
| | | $\mathbf{F}_8(x_8)$ | 8 | $\mathbf{F}_8(\cdot)$ | |
| | | $\mathbf{F}_9(x_9)$ | 9 | $\mathbf{F}_9(\cdot)$ | |
| | | | | | \vdots |

achieved very efficiently from OT extension

Alice

Bob

a

c

b

d

c

e

d

f

Alice

m bins

| | |
|---|-------------|
| a | 1 2 3 |
| b | 4 5 |
| c | 6 7 |
| d | 8 9 |
| | 10 |

Bob

c

d

e

f

1. Agree on random

$$h_1, h_2 : \{0, 1\}^* \rightarrow [m]$$

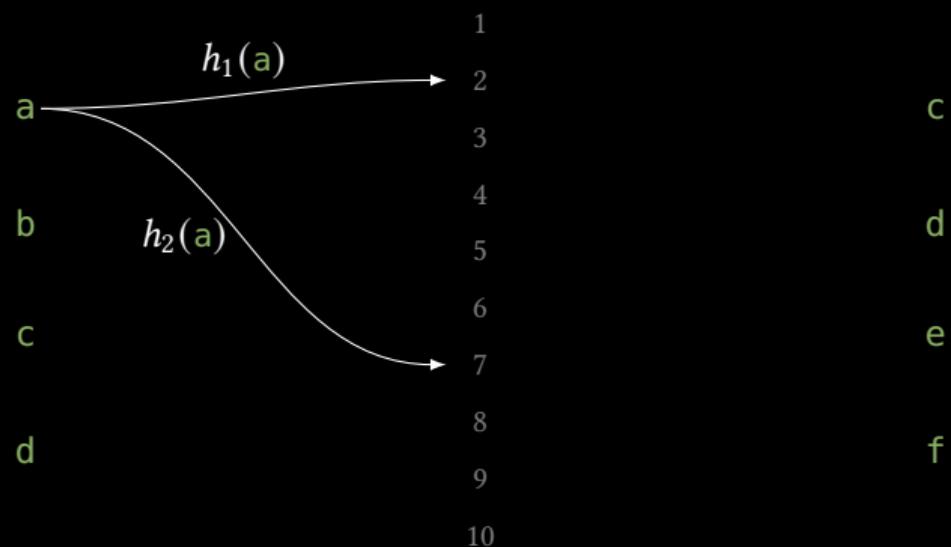
Alice

m bins

Bob

1. Agree on random

$$h_1, h_2 : \{0, 1\}^* \rightarrow [m]$$



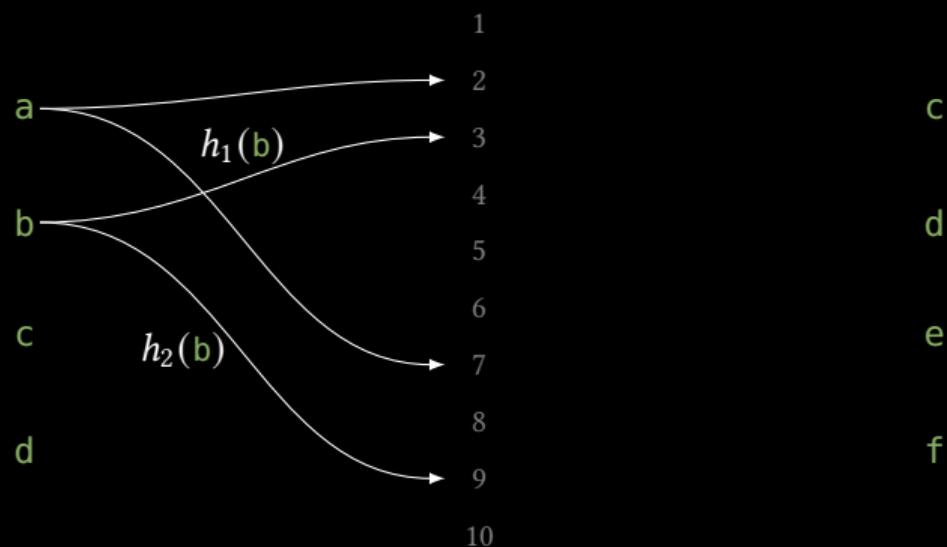
Alice

m bins

Bob

1. Agree on random

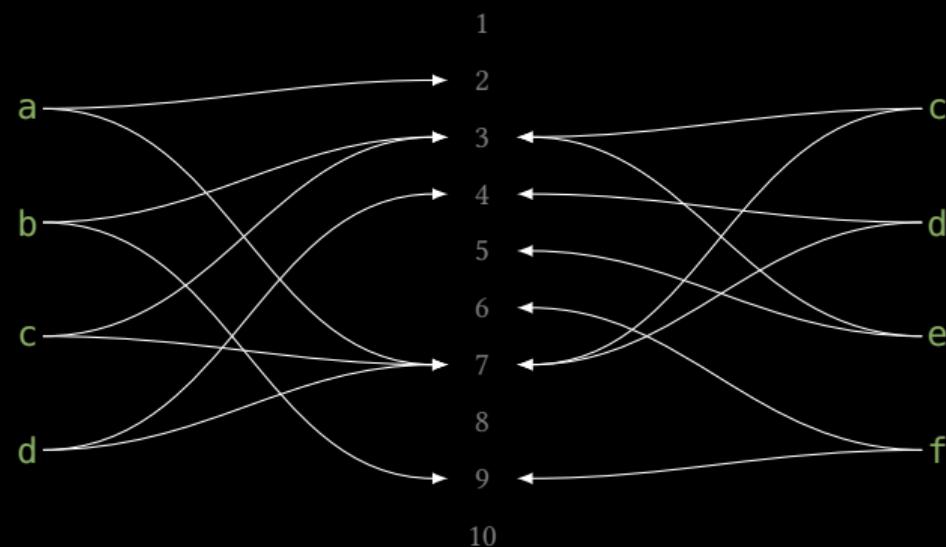
$$h_1, h_2 : \{0, 1\}^* \rightarrow [m]$$



Alice

m bins

Bob



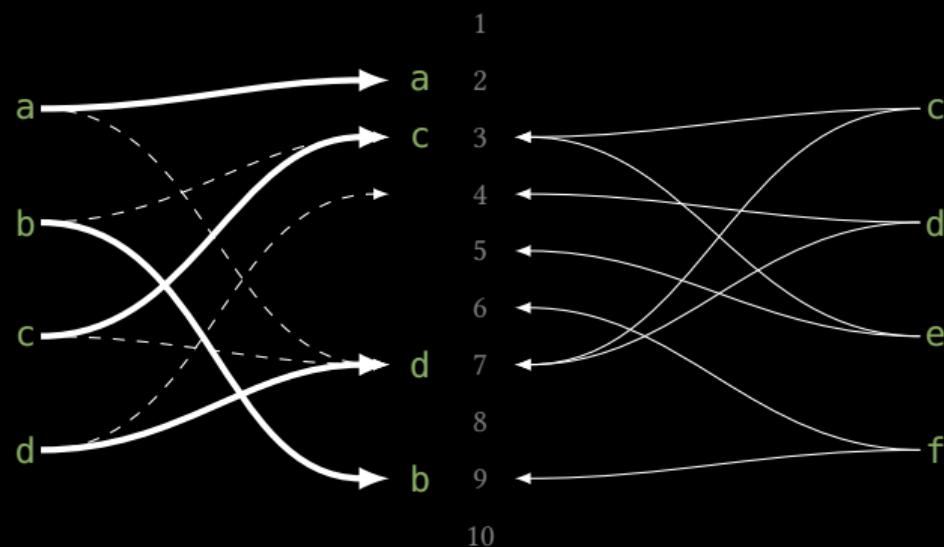
1. Agree on random

$$h_1, h_2 : \{0, 1\}^* \rightarrow [m]$$

Alice

m bins

Bob



1. Agree on random

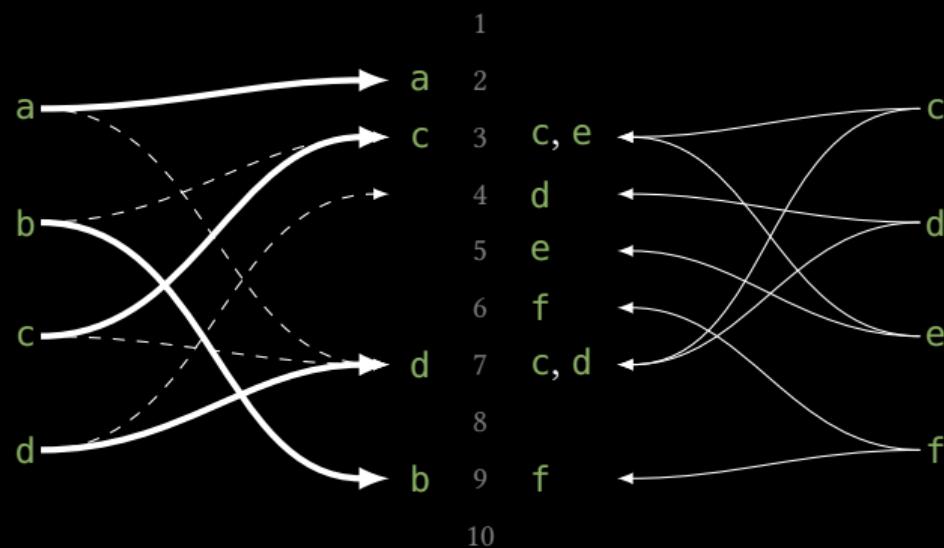
$$h_1, h_2 : \{0, 1\}^* \rightarrow [m]$$

2. Alice places each x into bin $h_1(x)$ or $h_2(x)$

Alice

m bins

Bob

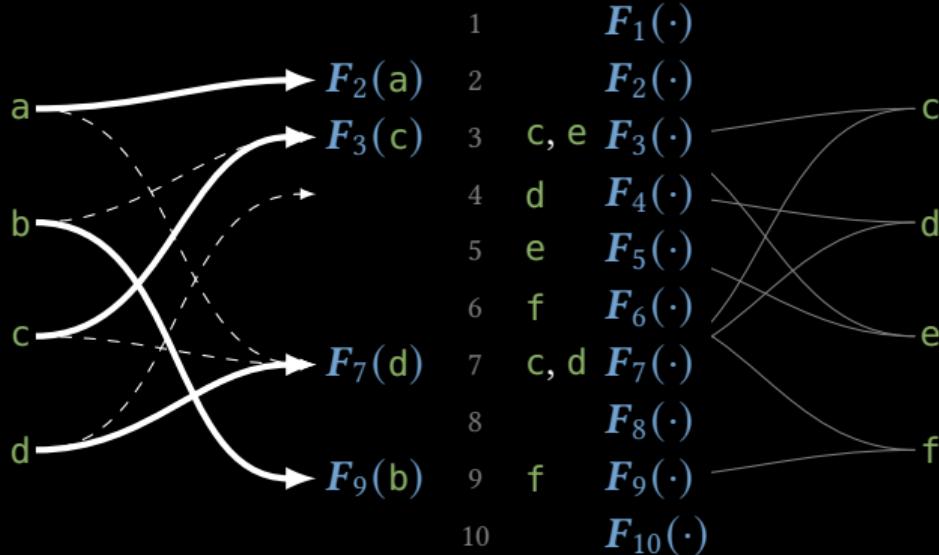


1. Agree on random
 $h_1, h_2 : \{0, 1\}^* \rightarrow [m]$
2. Alice places each x into bin $h_1(x)$ or $h_2(x)$
3. Bob places each x into bins $h_1(x)$ and $h_2(x)$

Alice

m bins

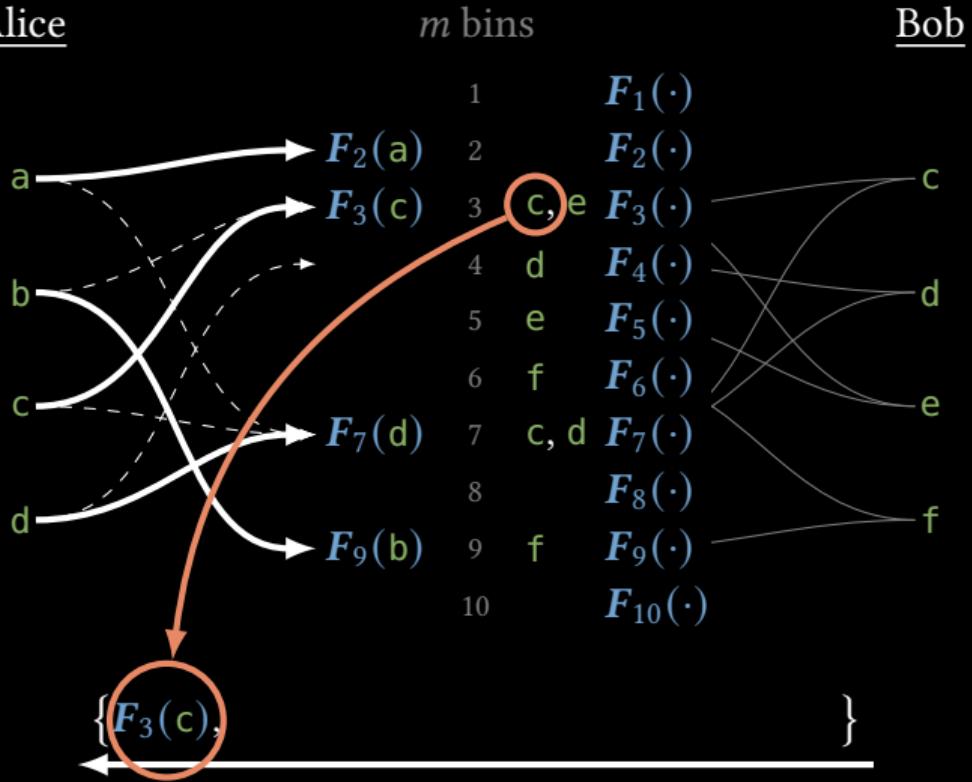
Bob



1. Agree on random
 $h_1, h_2 : \{0, 1\}^* \rightarrow [m]$
 2. Alice places each x into
bin $h_1(x)$ or $h_2(x)$
 3. Bob places each x into
bins $h_1(x)$ and $h_2(x)$
 4. OPRF in each bin:
Alice learns one $F_i(x)$;
Bob learns entire $F_i(\cdot)$

[PinkasSchneiderZohner14, KolesnikovKumaresanRosulekTrieu16]

Alice

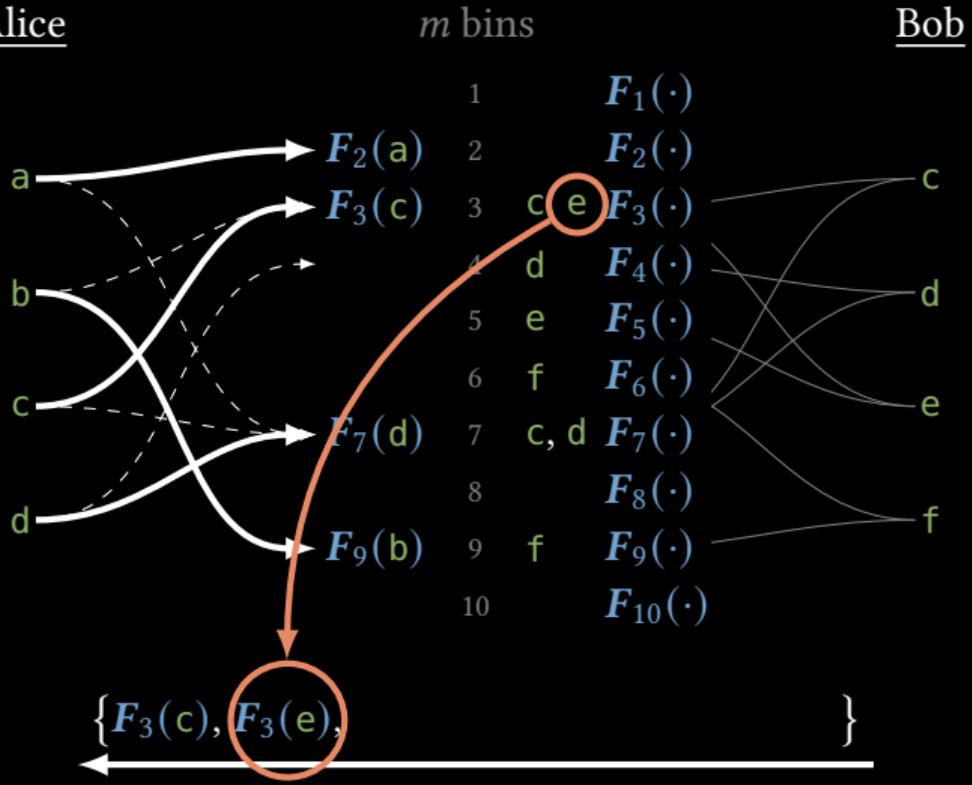


Bob

1. Agree on random $h_1, h_2 : \{0, 1\}^* \rightarrow [m]$
2. Alice places each x into bin $h_1(x)$ or $h_2(x)$
3. Bob places each x into bins $h_1(x)$ and $h_2(x)$
4. OPRF in each bin:
Alice learns one $F_i(x)$;
Bob learns entire $F_i(\cdot)$
5. Bob sends all $F_i(x)$ values

[PinkasSchneiderZohner14, KolesnikovKumaresanRosulekTrieu16]

Alice

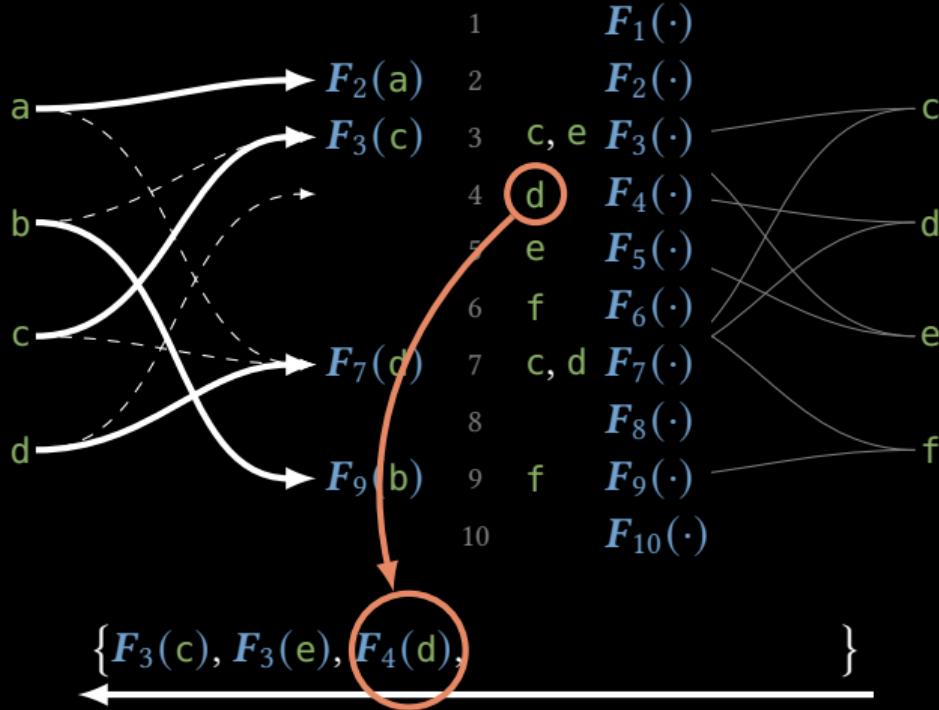


Bob

1. Agree on random $h_1, h_2 : \{0, 1\}^* \rightarrow [m]$
2. Alice places each x into bin $h_1(x)$ or $h_2(x)$
3. Bob places each x into bins $h_1(x)$ and $h_2(x)$
4. OPRF in each bin:
Alice learns one $F_i(x)$;
Bob learns entire $F_i(\cdot)$
5. Bob sends all $F_i(x)$ values

[PinkasSchneiderZohner14, KolesnikovKumaresanRosulekTrieu16]

Alice

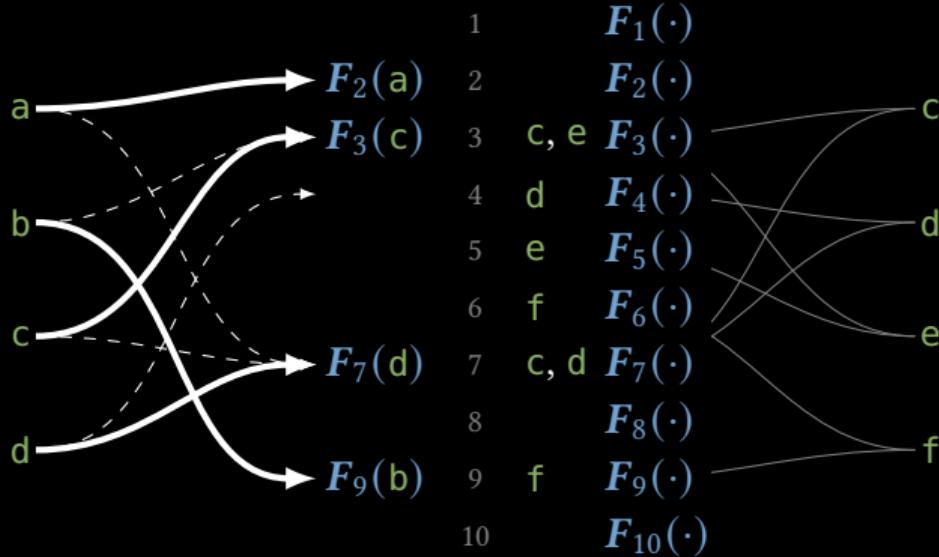


Bob

1. Agree on random $h_1, h_2 : \{0, 1\}^* \rightarrow [m]$
2. Alice places each x into bin $h_1(x)$ or $h_2(x)$
3. Bob places each x into bins $h_1(x)$ and $h_2(x)$
4. OPRF in each bin:
Alice learns one $F_i(x)$;
Bob learns entire $F_i(\cdot)$
5. Bob sends all $F_i(x)$ values

[PinkasSchneiderZohner14, KolesnikovKumaresanRosulekTrieu16]

Alice



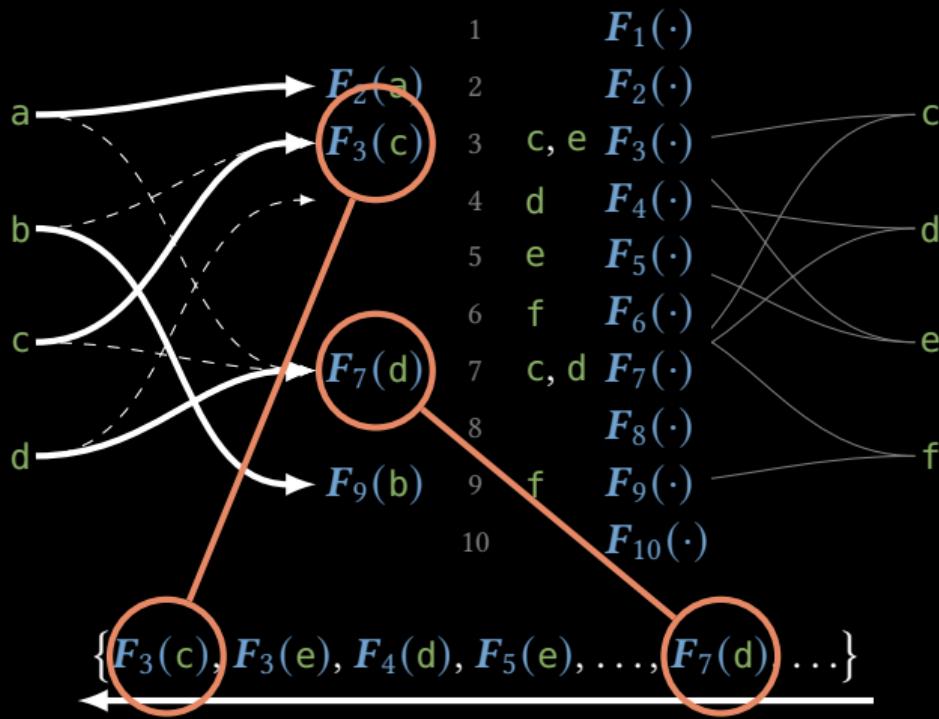
$$\{\mathbf{F}_3(\textcolor{brown}{c}), \mathbf{F}_3(\textcolor{brown}{e}), \mathbf{F}_4(\textcolor{brown}{d}), \mathbf{F}_5(\textcolor{brown}{e}), \dots, \mathbf{F}_7(\textcolor{brown}{d}), \dots\}$$

Bob

1. Agree on random $h_1, h_2 : \{0, 1\}^* \rightarrow [m]$
2. Alice places each x into bin $h_1(x)$ or $h_2(x)$
3. Bob places each x into bins $h_1(x)$ and $h_2(x)$
4. OPRF in each bin:
Alice learns one $\mathbf{F}_i(x)$;
Bob learns entire $\mathbf{F}_i(\cdot)$
5. Bob sends all $\mathbf{F}_i(x)$ values

[PinkasSchneiderZohner14, KolesnikovKumaresanRosulekTrieu16]

Alice



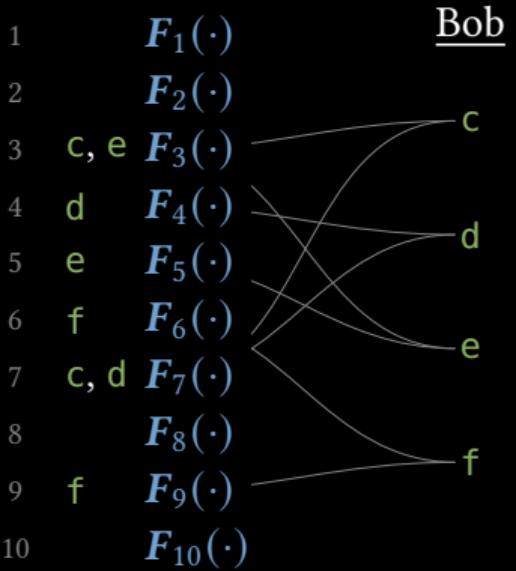
Bob

1. Agree on random $h_1, h_2 : \{0, 1\}^* \rightarrow [m]$
2. Alice places each x into bin $h_1(x)$ or $h_2(x)$
3. Bob places each x into bins $h_1(x)$ and $h_2(x)$
4. OPRF in each bin:
Alice learns one $F_i(x)$;
Bob learns entire $F_i(\cdot)$
5. Bob sends all $F_i(x)$ values

why isn't it secure against malicious parties?

why isn't it secure against malicious parties?

Alice

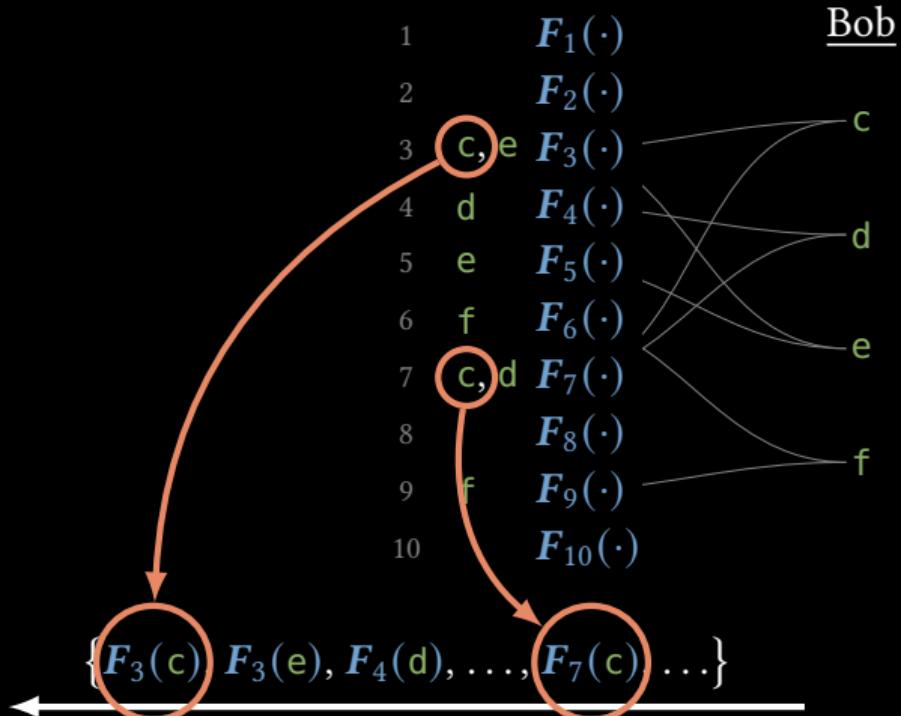


$$\{\mathbf{F}_3(\text{c}), \mathbf{F}_3(\text{e}), \mathbf{F}_4(\text{d}), \dots, \mathbf{F}_7(\text{c}), \dots\}$$



why isn't it secure against malicious parties?

Alice



Bob should send two
 F -values per item

why isn't it secure against malicious parties?

Alice

1 $F_1(\cdot)$

2 $F_2(\cdot)$

3 c, e $F_3(\cdot)$

4 d $F_4(\cdot)$

5 e $F_5(\cdot)$

6 f $F_6(\cdot)$

7 c, d $F_7(\cdot)$

8 $F_8(\cdot)$

9 f $F_9(\cdot)$

10 $F_{10}(\cdot)$

Bob

c

d

e

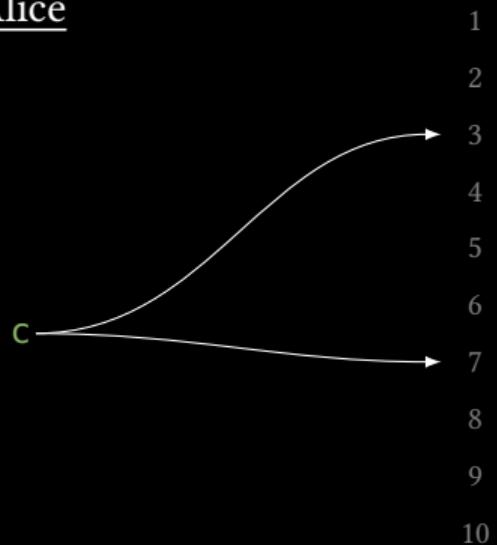
f

Bob should send two
 F -values per item , what if
he sends **only one**?

$\{F_3(c), F_3(e), F_4(d), \dots, \cancel{F_7(c)}, \dots\}$

why isn't it secure against malicious parties?

Alice



Bob

Bob should send two
F-values per item , what if
he sends **only one**?

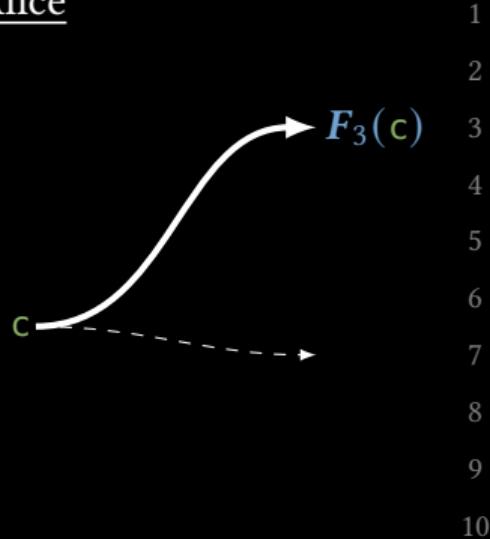
Alice has **c**; does she
include it in output?

$$\{F_3(c), F_3(e), F_4(d), \dots, \cancel{F_7(c)}, \dots\}$$



why isn't it secure against malicious parties?

Alice



Bob

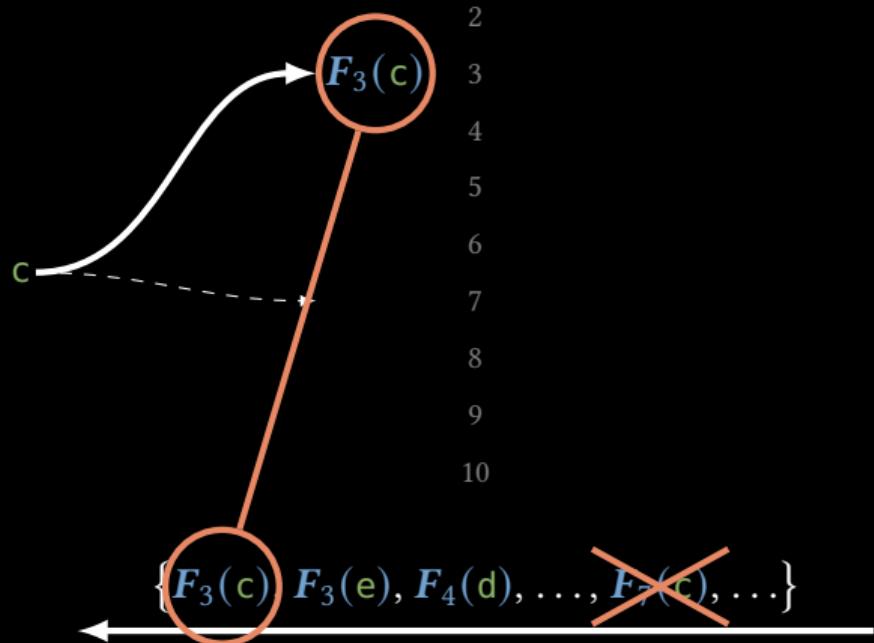
Bob should send two
 F -values per item , what if
he sends **only one**?

Alice has **c**; does she
include it in output?

$\{F_3(c), F_3(e), F_4(d), \dots, \cancel{F_7(c)}, \dots\}$

why isn't it secure against malicious parties?

Alice



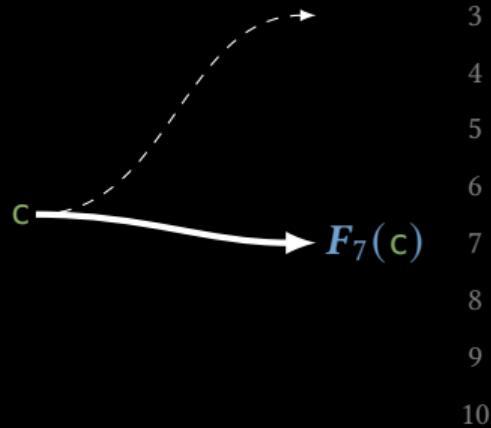
Bob

Bob should send two
 F -values per item , what if
he sends **only one**?

Alice has **c**; does she
include it in output?

why isn't it secure against malicious parties?

Alice



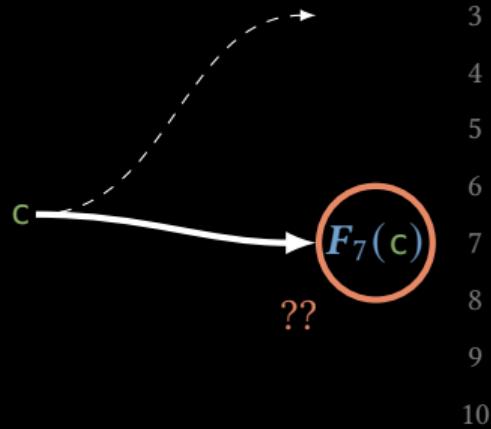
Bob

Bob should send two **F**-values per item , what if he sends **only one**?

Alice has **c**; does she include it in output?

why isn't it secure against malicious parties?

Alice



Bob

Bob should send two **F**-values per item , what if he sends **only one**?

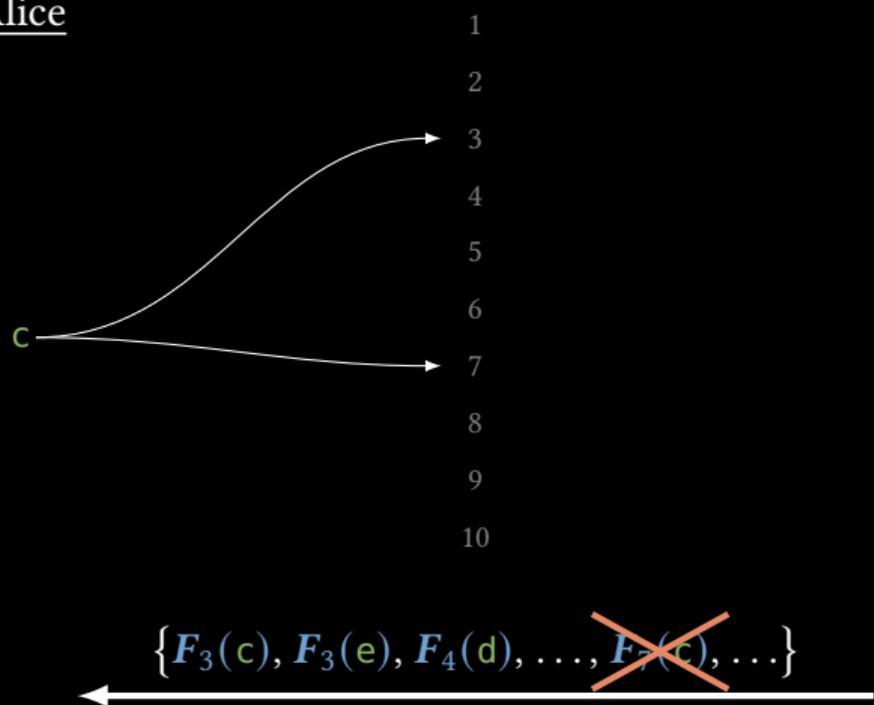
Alice has **c**; does she include it in output?

Only if **c** placed in bin 3!

$$\left\{ \mathbf{F}_3(\mathbf{c}), \mathbf{F}_3(\mathbf{e}), \mathbf{F}_4(\mathbf{d}), \dots, \cancel{\mathbf{F}_7(\mathbf{c})}, \dots \right\}$$

why isn't it secure against malicious parties?

Alice



Bob

Bob should send two **F**-values per item , what if he sends **only one**?

Alice has **c**; does she include it in output?

Only if **c** placed in bin 3!

- Depends on Alice's **entire input!**

⇒ can't simulate!

how do we overcome this problem?

[PinkasRosulekTrieuYanai20]

batch OPRF for malicious PSI

Alice Bob

$\mathbf{F}_1(x_1)$ 1 $\mathbf{F}_1(\cdot)$

$\mathbf{F}_2(x_2)$ 2 $\mathbf{F}_2(\cdot)$

$\mathbf{F}_3(x_3)$ 3 $\mathbf{F}_3(\cdot)$

$\mathbf{F}_4(x_4)$ 4 $\mathbf{F}_4(\cdot)$

$\mathbf{F}_5(x_5)$ 5 $\mathbf{F}_5(\cdot)$

$\mathbf{F}_6(x_6)$ 6 $\mathbf{F}_6(\cdot)$

$\mathbf{F}_7(x_7)$ 7 $\mathbf{F}_7(\cdot)$

$\mathbf{F}_8(x_8)$ 8 $\mathbf{F}_8(\cdot)$

$\mathbf{F}_9(x_9)$ 9 $\mathbf{F}_9(\cdot)$

⋮

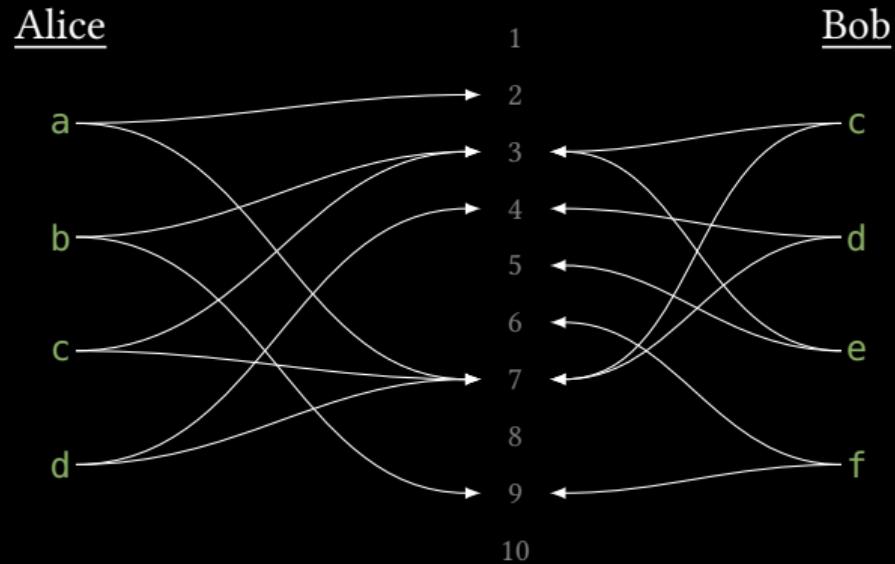
batch OPRF for malicious PSI

| <u>Alice</u> | | <u>Bob</u> | |
|--------------|---|--------------|--|
| $F_1(x_1)$ | 1 | $F_1(\cdot)$ | State of the art malicious batch OPRF [OrrùOrsiniScholl17] |
| $F_2(x_2)$ | 2 | $F_2(\cdot)$ | ► essentially same cost as semi-honest |
| $F_3(x_3)$ | 3 | $F_3(\cdot)$ | |
| $F_4(x_4)$ | 4 | $F_4(\cdot)$ | |
| $F_5(x_5)$ | 5 | $F_5(\cdot)$ | |
| $F_6(x_6)$ | 6 | $F_6(\cdot)$ | |
| $F_7(x_7)$ | 7 | $F_7(\cdot)$ | |
| $F_8(x_8)$ | 8 | $F_8(\cdot)$ | |
| $F_9(x_9)$ | 9 | $F_9(\cdot)$ | |
| | ⋮ | | |

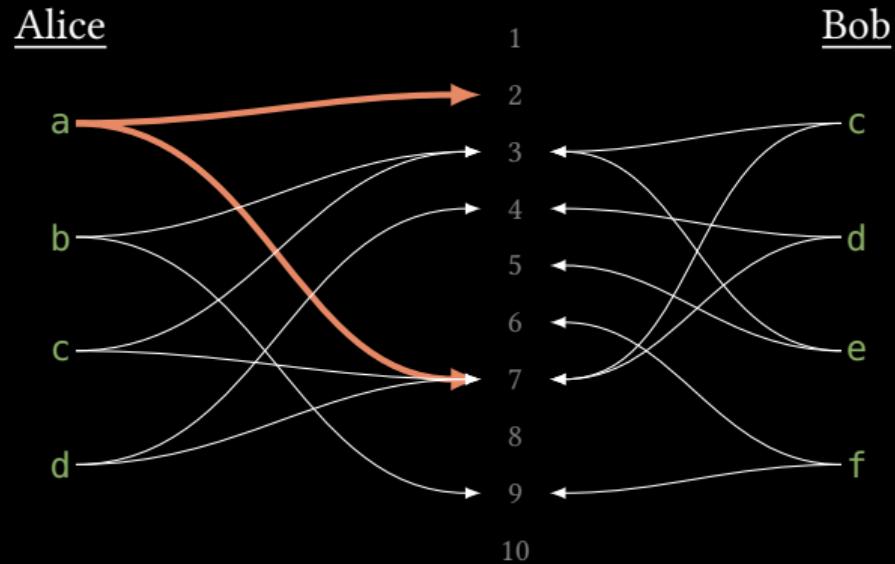
batch OPRF for malicious PSI

| <u>Alice</u> | | <u>Bob</u> | |
|--------------|---|--------------|--|
| $F_1(x_1)$ | 1 | $F_1(\cdot)$ | State of the art malicious batch OPRF [OrrùOrsiniScholl17] |
| $F_2(x_2)$ | 2 | $F_2(\cdot)$ | ► essentially same cost as semi-honest |
| $F_3(x_3)$ | 3 | $F_3(\cdot)$ | ► consistency check relies on an additive homomorphism: |
| $F_4(x_4)$ | 4 | $F_4(\cdot)$ | |
| $F_5(x_5)$ | 5 | $F_5(\cdot)$ | $F_i(x) \oplus F_j(y) = F_{ij}(x \oplus y)$ |
| $F_6(x_6)$ | 6 | $F_6(\cdot)$ | |
| $F_7(x_7)$ | 7 | $F_7(\cdot)$ | |
| $F_8(x_8)$ | 8 | $F_8(\cdot)$ | |
| $F_9(x_9)$ | 9 | $F_9(\cdot)$ | |
| | | \vdots | $^*: a gross oversimplification$ |

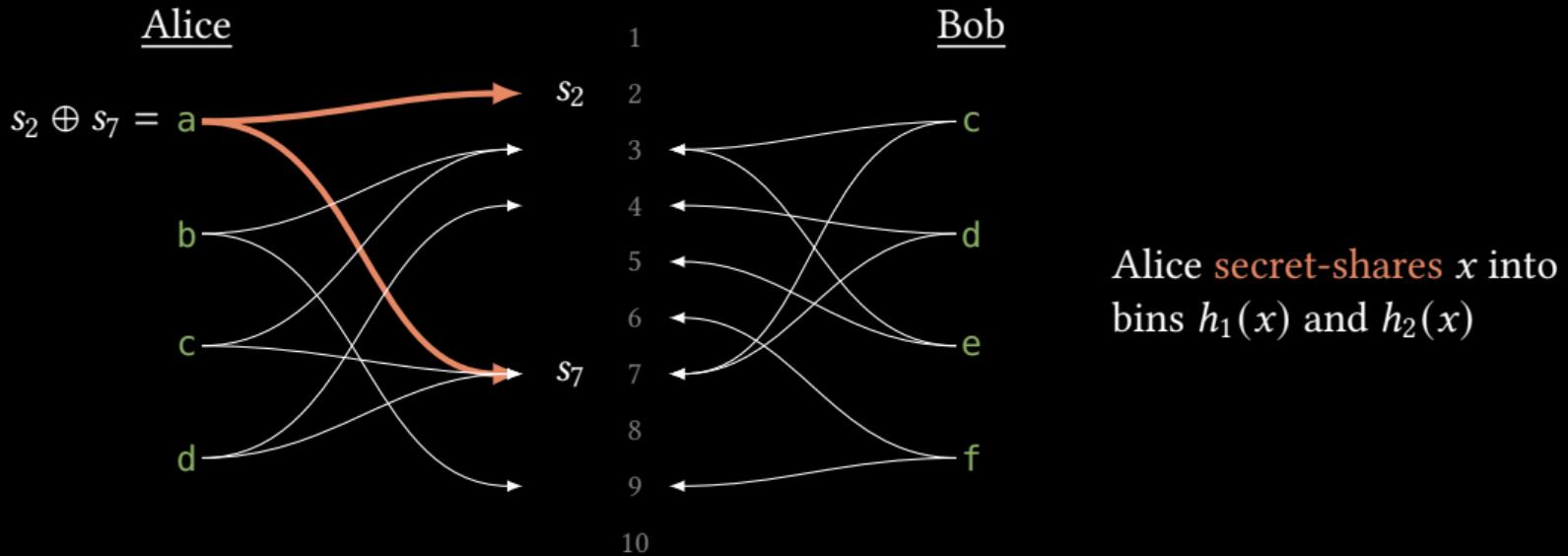
[PinkasRosulekTrieuYanai20] *protocol main idea:*



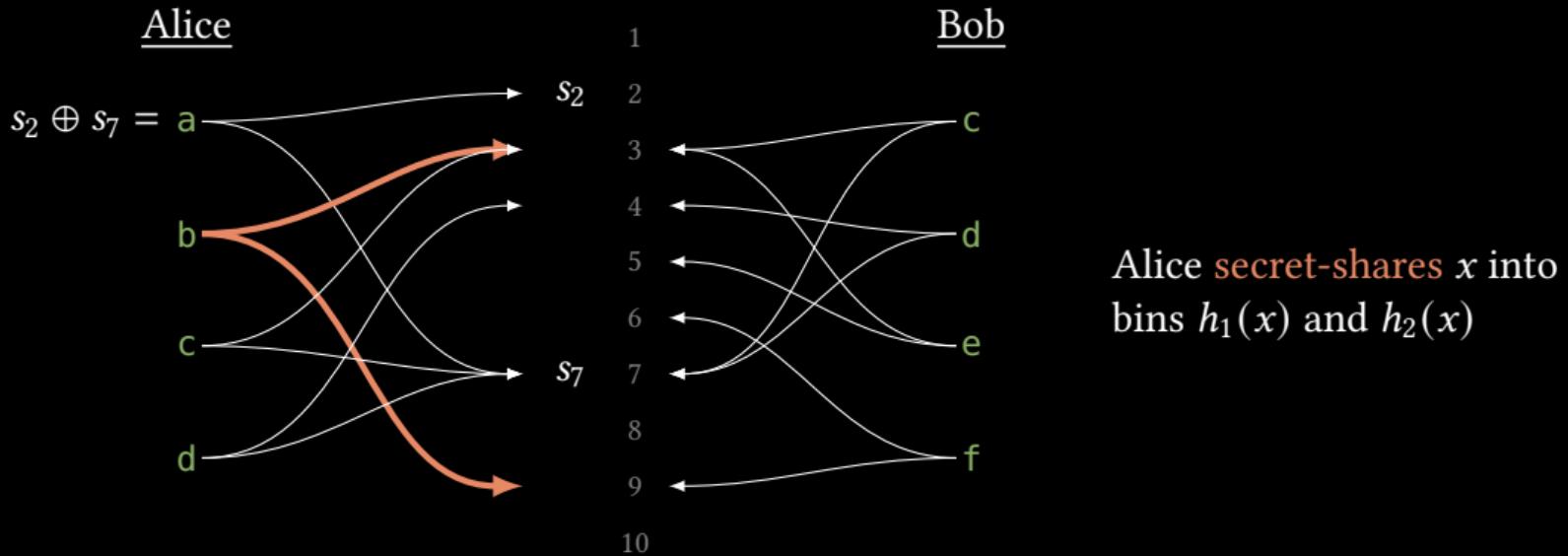
[PinkasRosulekTrieuYanai20] *protocol main idea:*



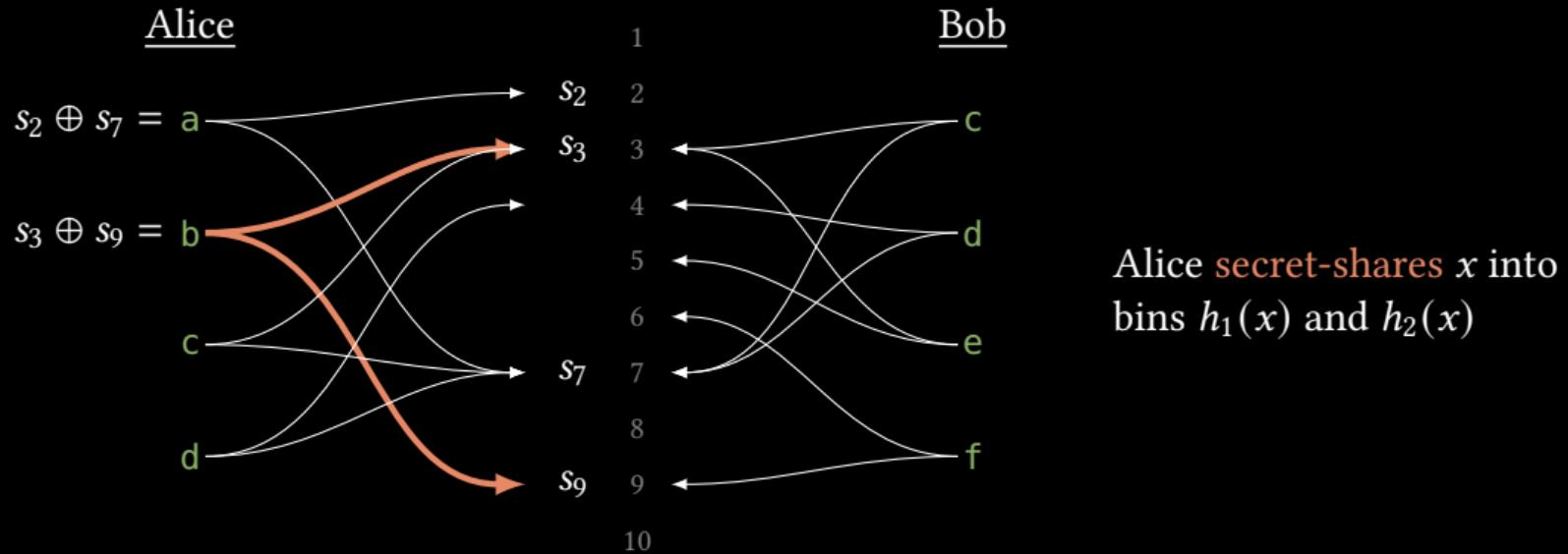
[PinkasRosulekTrieuYanai20] *protocol main idea:*



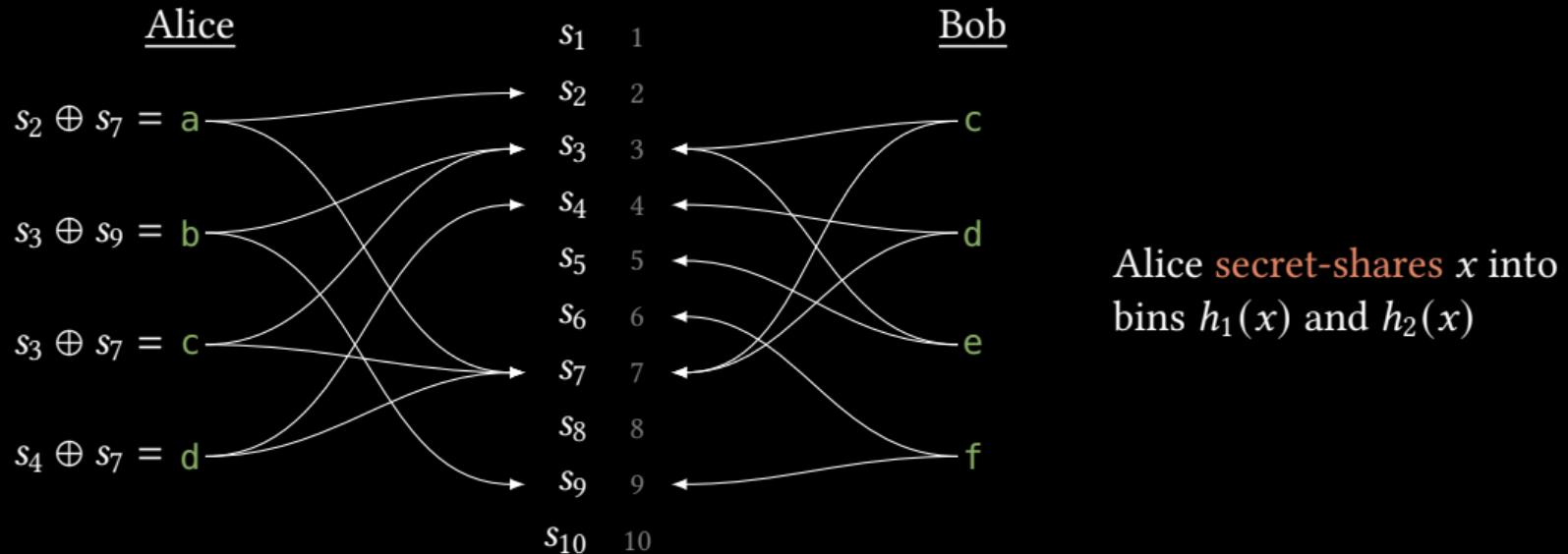
[PinkasRosulekTrieuYanai20] *protocol main idea:*



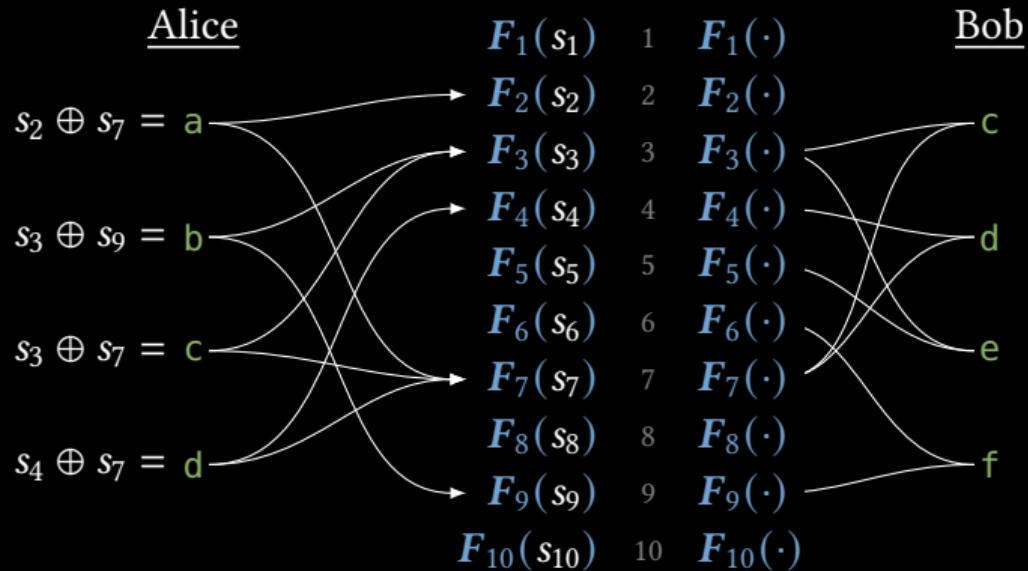
[PinkasRosulekTrieuYanai20] *protocol main idea:*



[PinkasRosulekTrieuYanai20] *protocol main idea:*

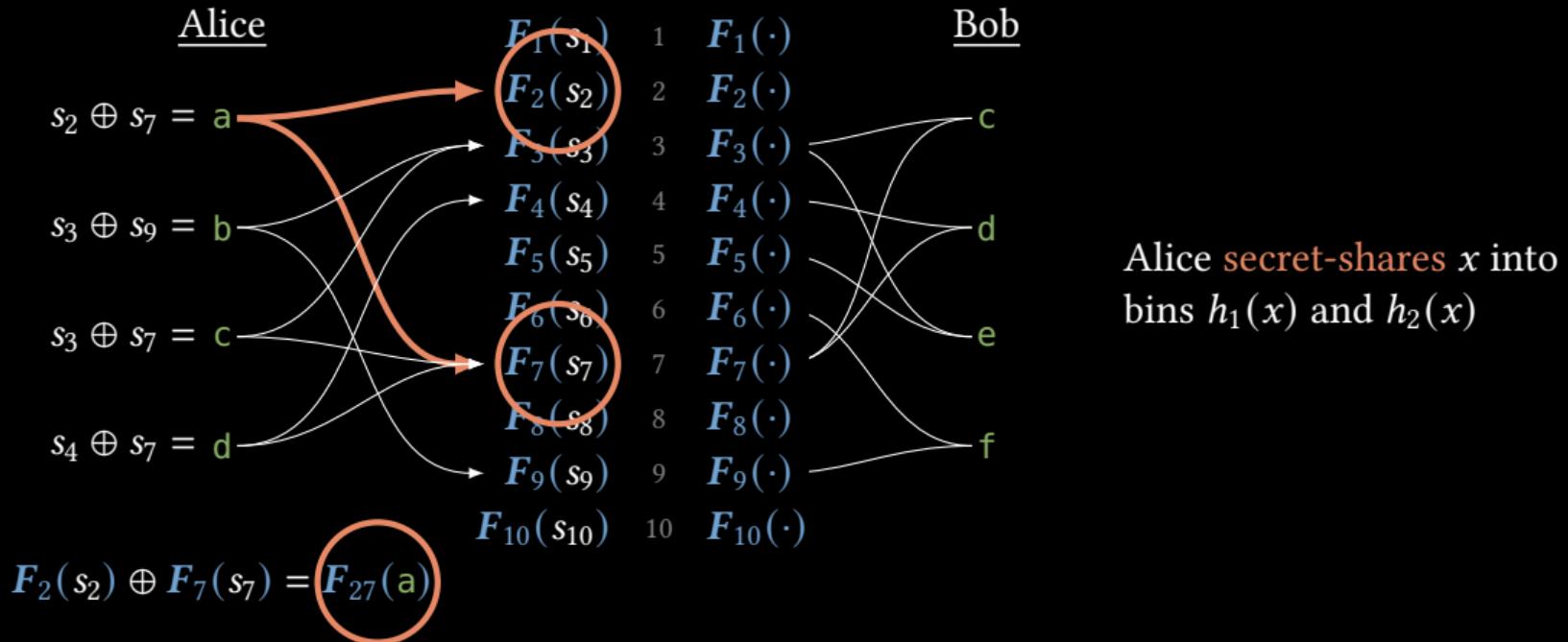


[PinkasRosulekTrieuYanai20] *protocol main idea:*

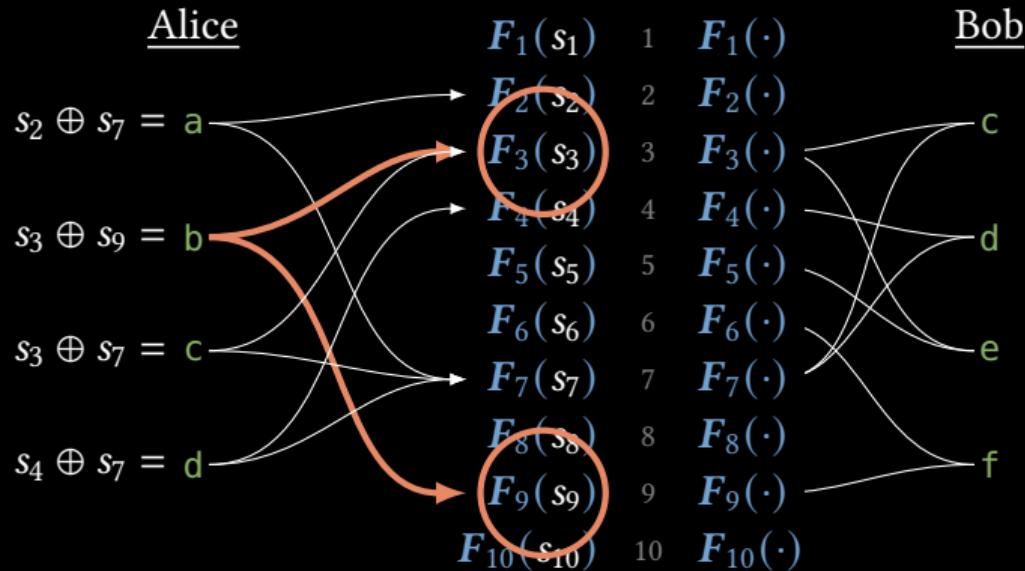


Alice secret-shares x into bins $h_1(x)$ and $h_2(x)$

[PinkasRosulekTrieuYanai20] protocol main idea:



[PinkasRosulekTrieuYanai20] protocol main idea:

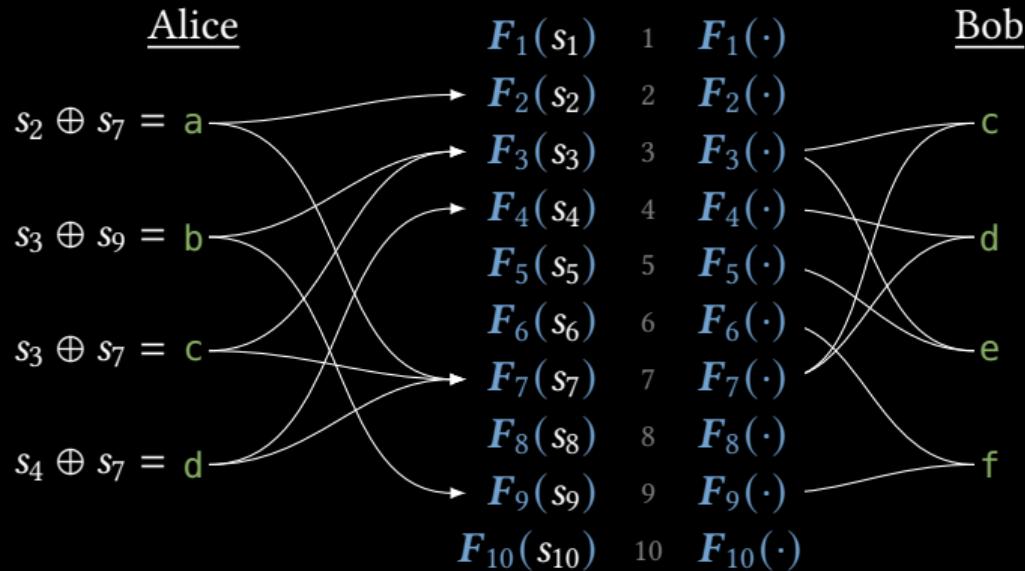


$$F_2(s_2) \oplus F_7(s_7) = F_{27}(a)$$

$$F_3(s_3) \oplus F_9(s_9) = F_{39}(b)$$

Alice secret-shares x into bins $h_1(x)$ and $h_2(x)$

[PinkasRosulekTrieuYanai20] *protocol main idea:*



$$F_2(s_2) \oplus F_7(s_7) = F_{27}(\text{a})$$

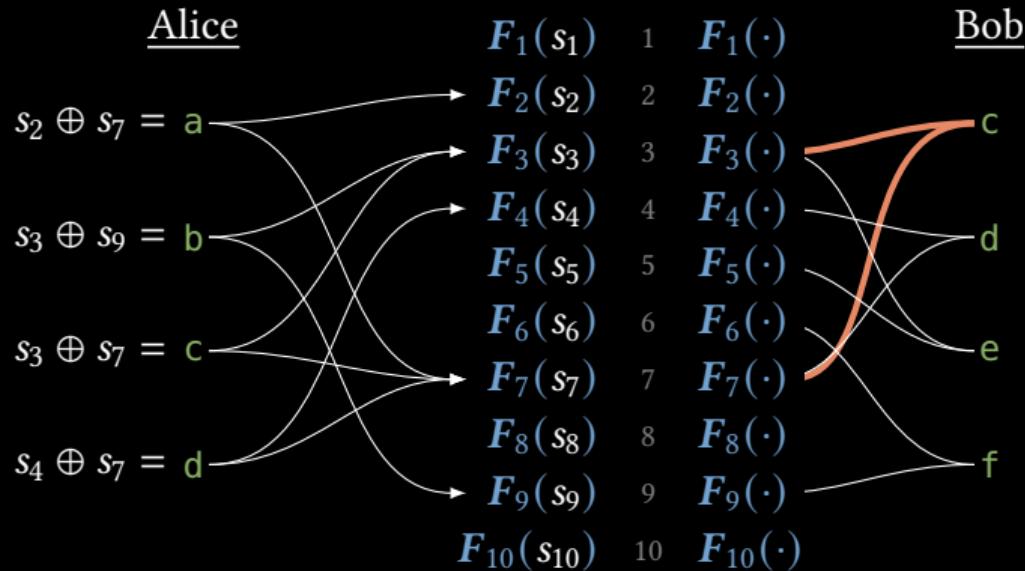
$$F_3(s_3) \oplus F_9(s_9) = F_{39}(\text{b})$$

$$F_3(s_3) \oplus F_7(s_7) = F_{37}(\text{c})$$

$$F_4(s_4) \oplus F_7(s_7) = F_{47}(\text{d})$$

Alice **secret-shares** x into bins $h_1(x)$ and $h_2(x)$

[PinkasRosulekTrieuYanai20] *protocol main idea:*



$$F_2(s_2) \oplus F_7(s_7) = F_{27}(a)$$

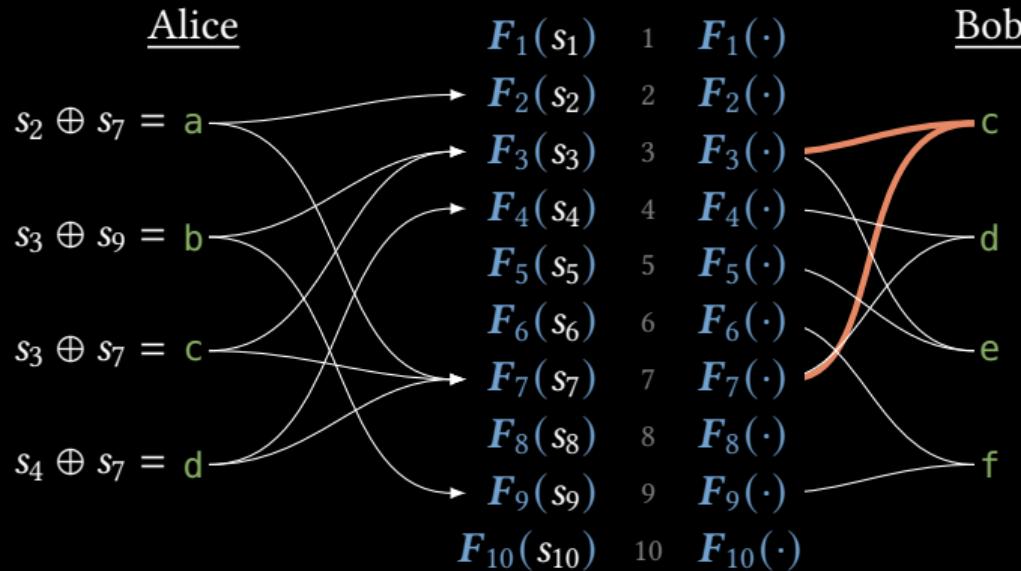
$$F_3(s_3) \oplus F_9(s_9) = F_{39}(b)$$

$$F_3(s_3) \oplus F_7(s_7) = F_{37}(c)$$

$$F_4(s_4) \oplus F_7(s_7) = F_{47}(d)$$

Alice secret-shares x into bins $h_1(x)$ and $h_2(x)$

[PinkasRosulekTrieuYanai20] *protocol main idea:*



$$F_2(s_2) \oplus F_7(s_7) = F_{27}(a)$$

$$F_3(s_3) \oplus F_9(s_9) = F_{39}(b)$$

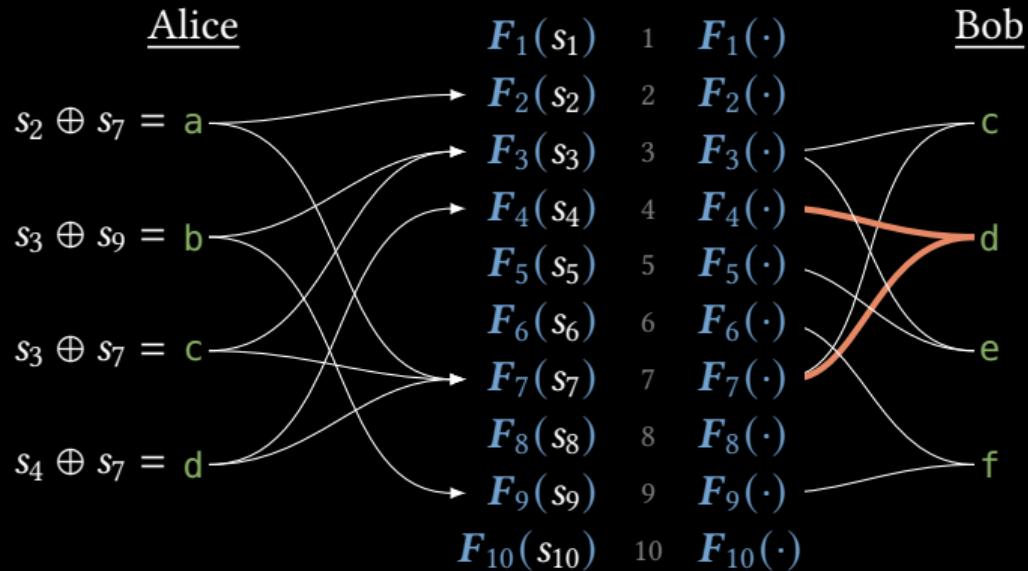
$$F_3(s_3) \oplus F_7(s_7) = F_{37}(c)$$

$$F_4(s_4) \oplus F_7(s_7) = F_{47}(d)$$

$$\{F_{37}(c), \quad \}$$

Alice secret-shares x into bins $h_1(x)$ and $h_2(x)$

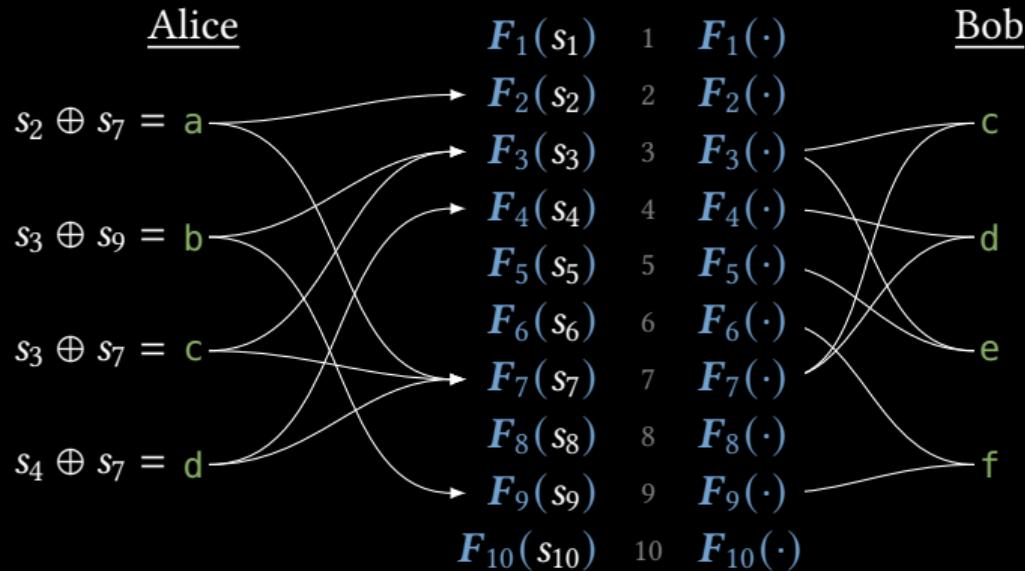
[PinkasRosulekTrieuYanai20] *protocol main idea:*



$$\begin{aligned}
 F_2(s_2) \oplus F_7(s_7) &= F_{27}(a) \\
 F_3(s_3) \oplus F_9(s_9) &= F_{39}(b) \\
 F_3(s_3) \oplus F_7(s_7) &= F_{37}(c) \\
 F_4(s_4) \oplus F_7(s_7) &= F_{47}(d)
 \end{aligned}
 \quad \left\{ F_{37}(c), \textcircled{F_{47}(d)}, \dots \right\}$$

Alice secret-shares x into bins $h_1(x)$ and $h_2(x)$

[PinkasRosulekTrieuYanai20] *protocol main idea:*



$$F_2(s_2) \oplus F_7(s_7) = F_{27}(a)$$

$$F_3(s_3) \oplus F_9(s_9) = F_{39}(b)$$

$$F_3(s_3) \oplus F_7(s_7) = F_{37}(c)$$

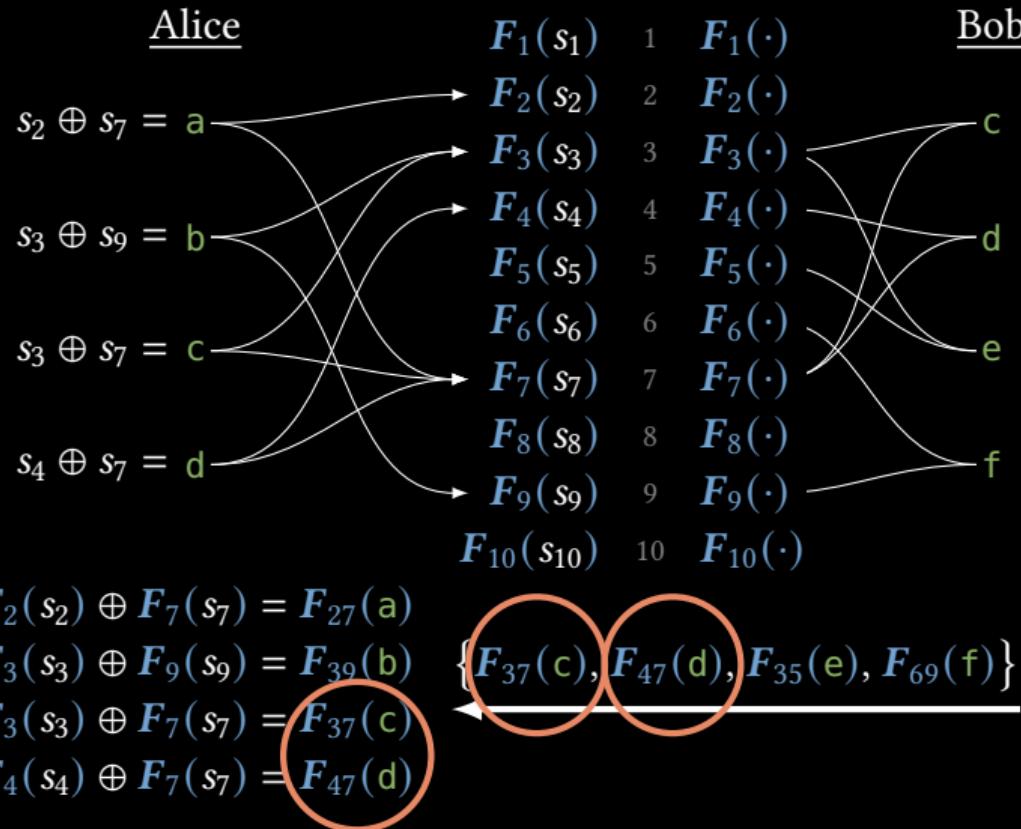
$$F_4(s_4) \oplus F_7(s_7) = F_{47}(d)$$

$$\{F_{37}(c), F_{47}(d), F_{35}(e), F_{69}(f)\}$$

Alice secret-shares x into bins $h_1(x)$ and $h_2(x)$

Bob sends only one F -value per item

[PinkasRosulekTrieuYanai20] *protocol main idea:*



Alice secret-shares x into bins $h_1(x)$ and $h_2(x)$

Bob sends only one F -value per item

overview: PSI on large sets

for 1 million items:

4.5 – 5 seconds; 128 – 145 MB

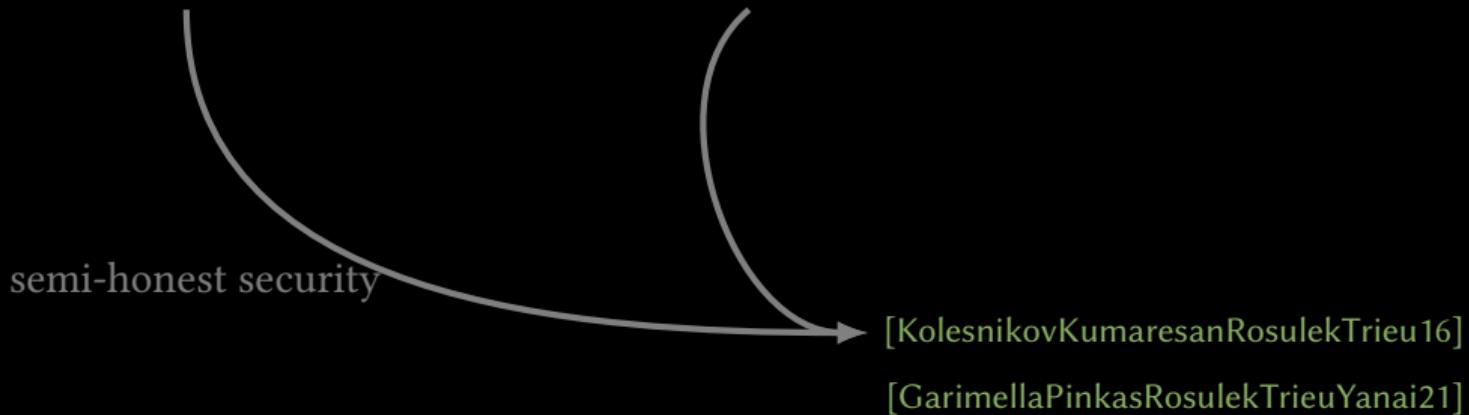
[KolesnikovKumaresanRosulekTrieu16]

[GirimellaPinkasRosulekTrieuYanai21]

overview: PSI on large sets

for 1 million items:

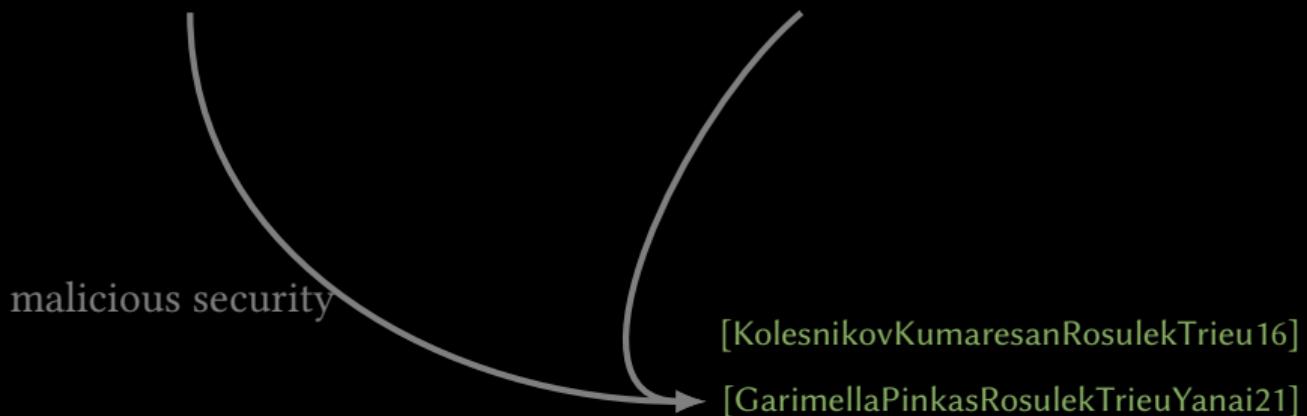
4.5 – 5 seconds; 128 – 145 MB



overview: PSI on large sets

for 1 million items:

4.5 – 5 seconds; 128 – 145 MB



*PSI on **asymmetric** sets*

$s_2 \otimes s_1 = d$
 $s_3 \otimes s_9 = b$

offline preprocessing techniques and leakage; scaling to billions of items

how to scale to billions of items?

The image shows a screenshot of a news article from The Verge. At the top, there's a navigation bar with the site's logo 'THE VERGE' and categories: TECH ▾, REVIEWS ▾, SCIENCE ▾, CREATORS ▾, and ENTERTAINMENT ▾. Below the navigation bar is a large, dark rectangular area. In the center of this area, there's a horizontal line with three colored segments: blue for 'APPS', black for 'TECH', and red for 'FACEBOOK'. Below this line, the main title of the article is displayed in a large, bold, black font: 'WhatsApp now has 2 billion users'. Underneath the title, a subtitle in a smaller, italicized black font reads: 'And it has no plans to drop end-to-end encryption'. At the bottom left of the dark area, the author information is provided: 'By Jon Porter | @JonPorter | Feb 12, 2020, 10:50am EST'.

THE VERGE

TECH ▾ REVIEWS ▾ SCIENCE ▾ CREATORS ▾ ENTERTAINMENT ▾

APPS \ TECH \ FACEBOOK

WhatsApp now has 2 billion users

And it has no plans to drop end-to-end encryption

By [Jon Porter](#) | [@JonPorter](#) | Feb 12, 2020, 10:50am EST

how to scale to billions of items?

The image is a collage of three screenshots. On the left, a news article from The Verge titled "WhatsApp n... And it has no plans to drop end-to-end encryption" by Jon Porter (@JonPorty) on Feb 12, 2016. The article discusses WhatsApp's commitment to end-to-end encryption. On the right, a screenshot of the Have I Been Pwned? website shows a search bar with the placeholder 'email address' and a 'pwned?' button. Below the search bar, there's a promotional message for 1Password: 'Generate secure, unique passwords for every account' with a link 'Learn more at 1Password.com'. At the bottom, there are four large numbers representing breached data: 493 pwned websites, 10,467,311,280 pwned accounts, 113,841 pastes, and 195,045,089 paste accounts.

THE VERGE TECH

APP TECH FACEBOOK

WhatsApp n...

And it has no plans to drop end-to-end encryption

By Jon Porter | @JonPorty | Feb 12, 2016

493 pwned websites

10,467,311,280 pwned accounts

113,841 pastes

195,045,089 paste accounts

';--have i been pwned?

Check if you have an account that has been compromised in a data breach

email address

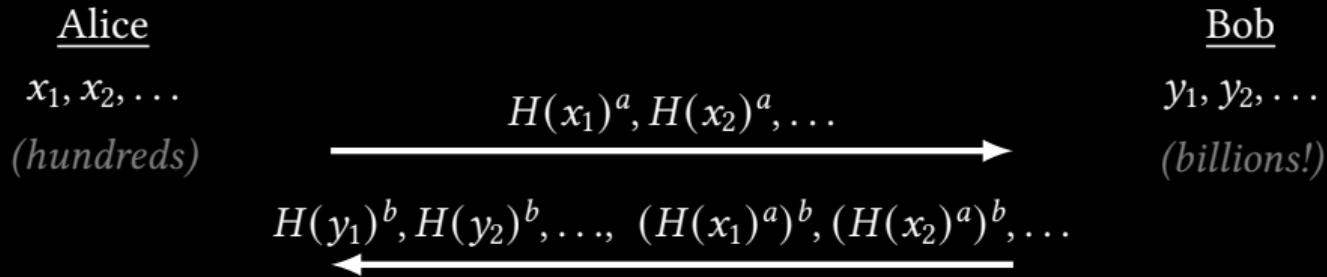
pwned?

Generate secure, unique passwords for every account

Why 1Password?

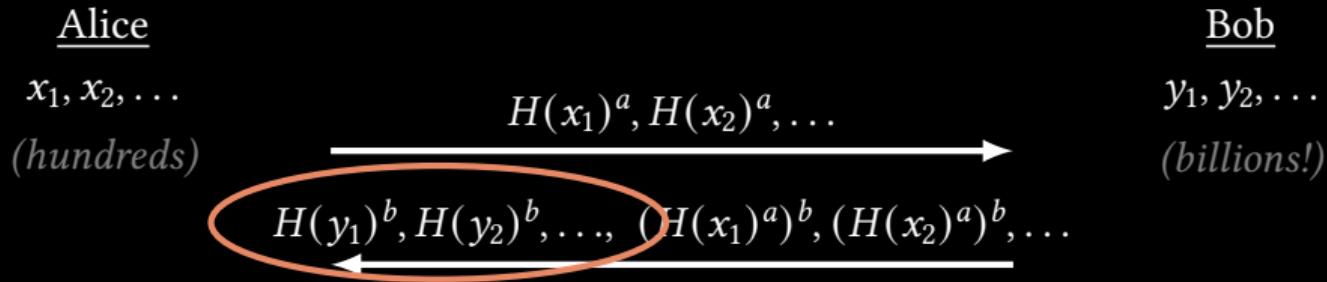
Learn more at 1Password.com

idea #1: offline preprocessing



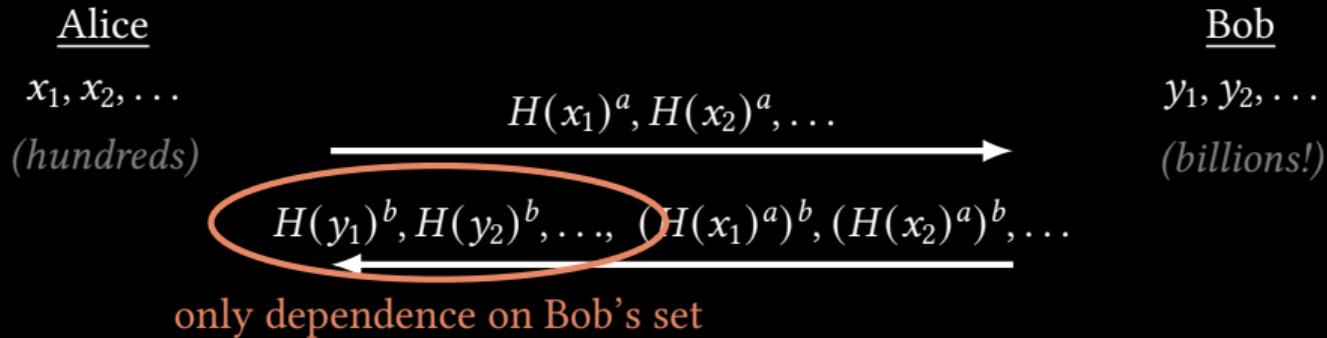
see [KalesRechbergerSchneiderSenkerWeinert19]

idea #1: offline preprocessing



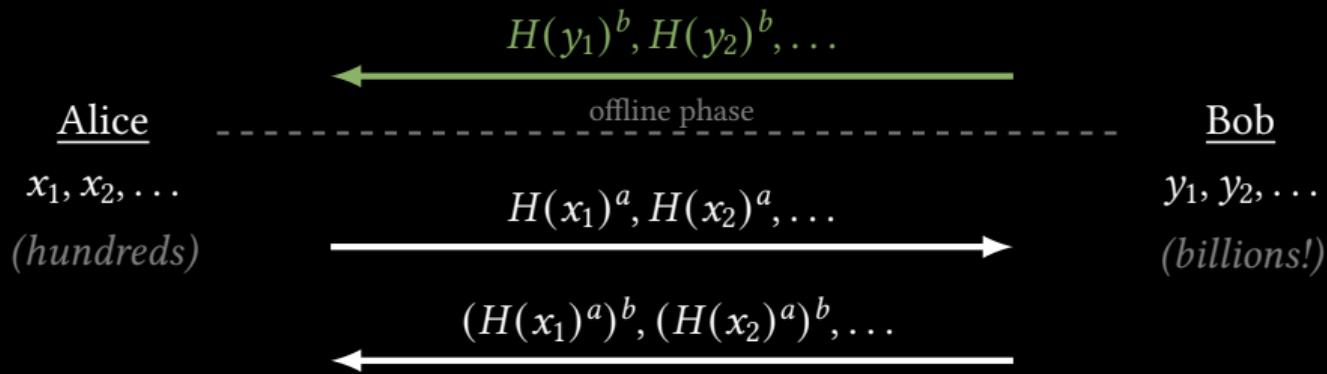
see [KalesRechbergerSchneiderSenkerWeinert19]

idea #1: offline preprocessing



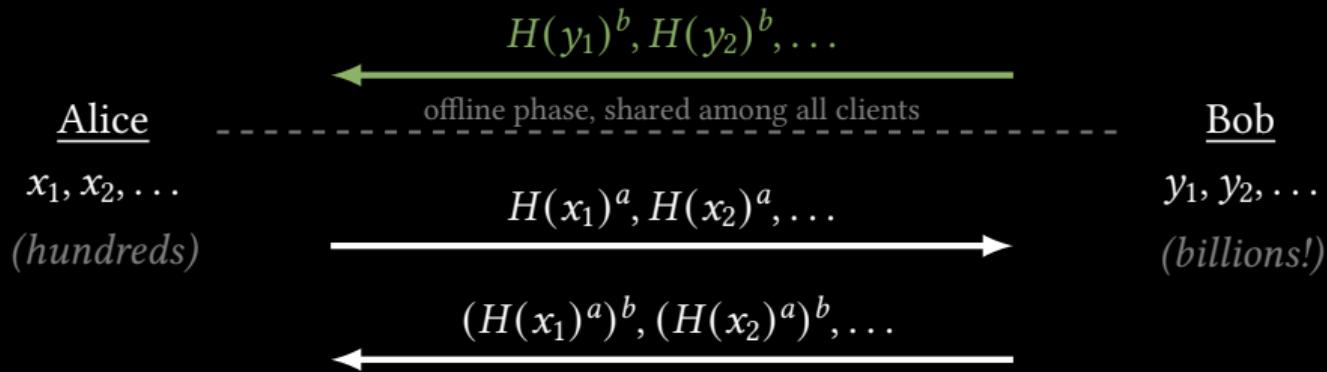
see [KalesRechbergerSchneiderSenkerWeinert19]

idea #1: offline preprocessing



see [KalesRechbergerSchneiderSenkerWeinert19]

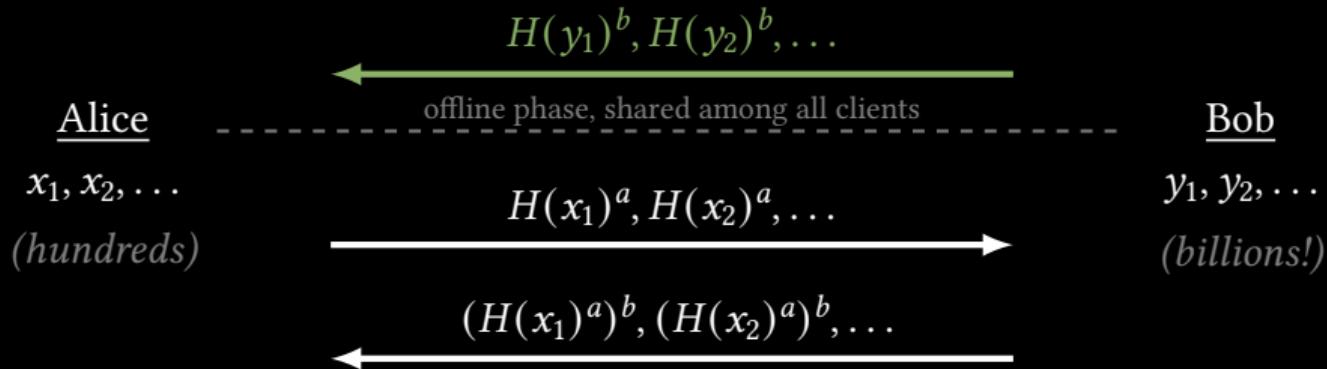
idea #1: offline preprocessing



- ▶ Safe to reuse b for many PSIs \Rightarrow reuse offline phase for all clients!

see [KalesRechbergerSchneiderSenkerWeinert19]

idea #1: offline preprocessing



- ▶ Safe to reuse b for many PSIs \Rightarrow reuse offline phase for all clients!
- ▶ Clever encodings for offline message: 4GB / 1B items

see [KalesRechbergerSchneiderSenkerWeinert19]

idea #2: allow some leakage

 100 items

Alice:

Bob:  1 billion items

see [LiPalAliSullivanChatterjeeRistenpart19]

idea #2: allow some leakage

100 items


Alice:

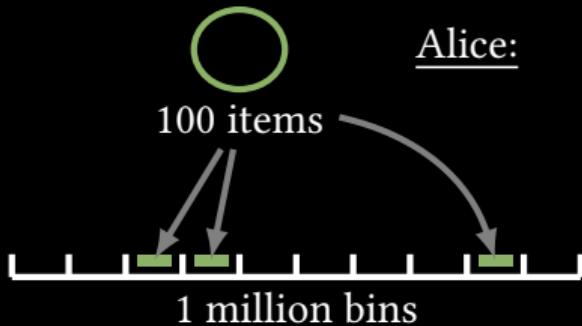


Bob: 
1 billion items

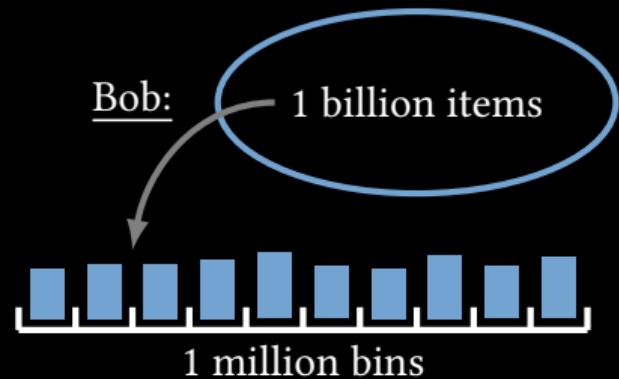


see [LiPalAliSullivanChatterjeeRistenpart19]

idea #2: allow some leakage

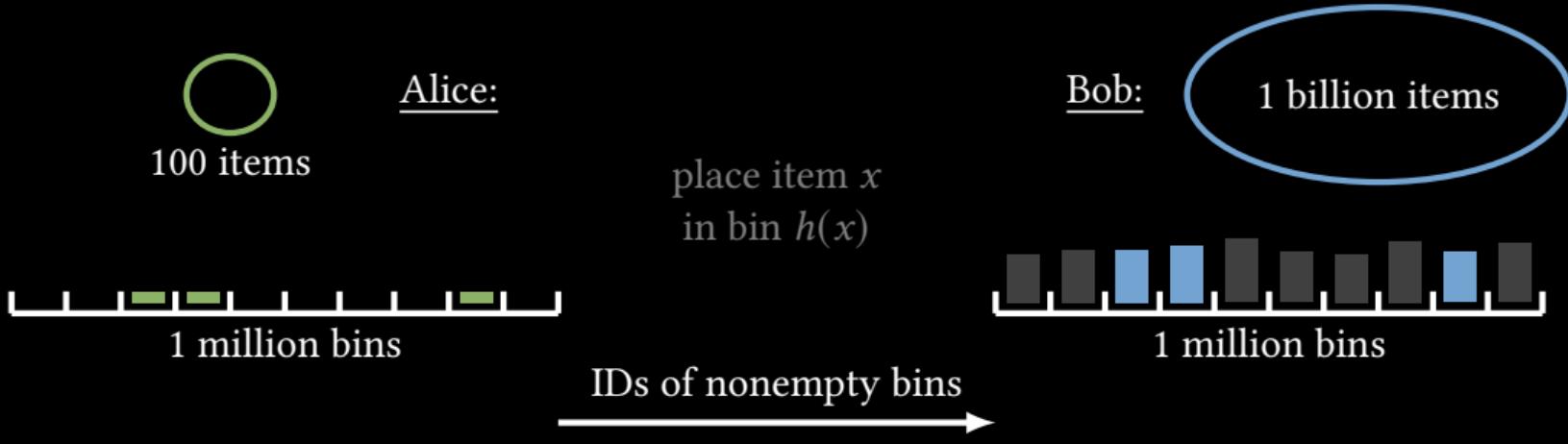


place item x
in bin $h(x)$



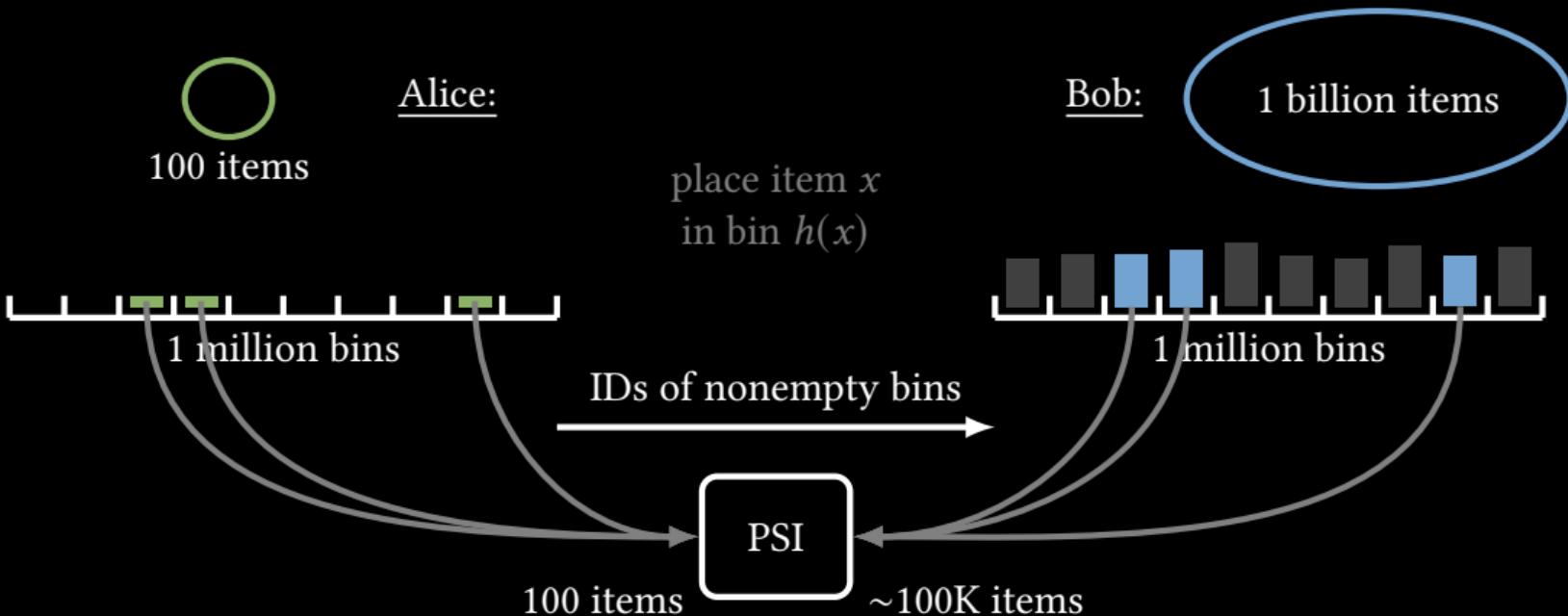
see [LiPalAliSullivanChatterjeeRistenpart19]

idea #2: allow some leakage



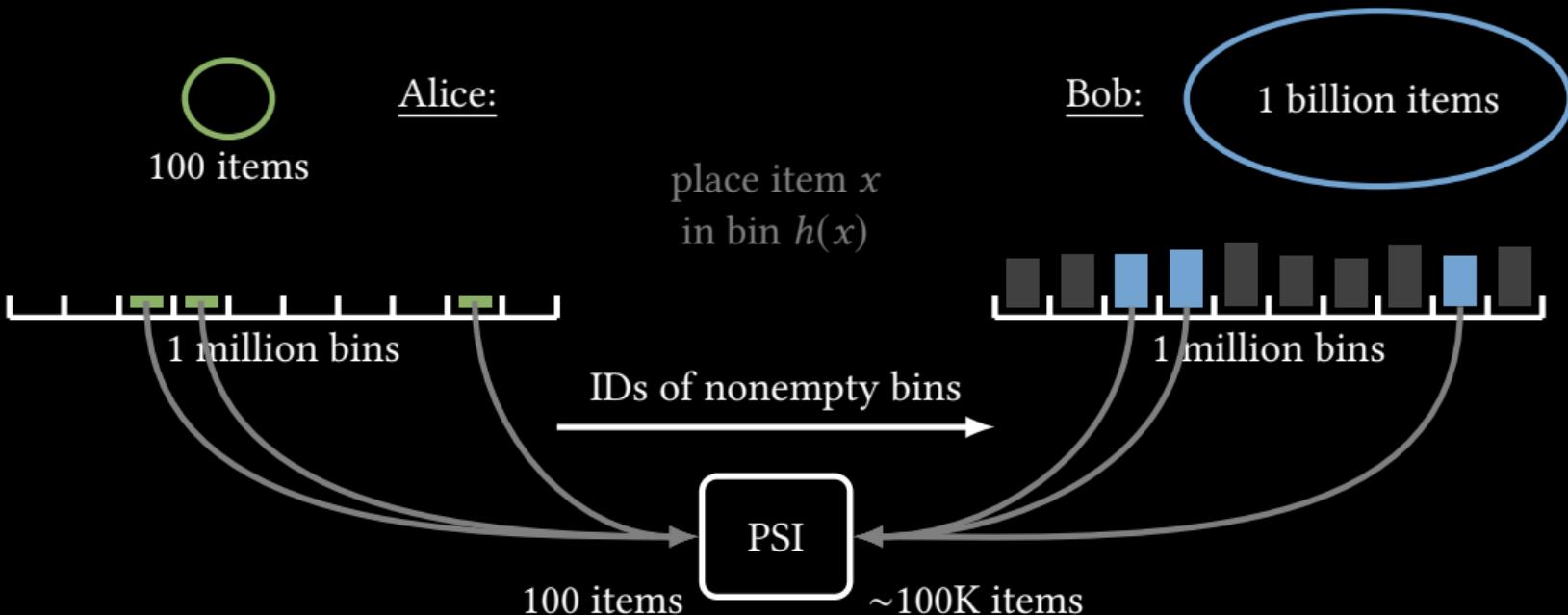
see [LiPalAliSullivanChatterjeeRistenpart19]

idea #2: allow some leakage



see [LiPalAliSullivanChatterjeeRistenpart19]

idea #2: allow some leakage



choice of h ? see [LiPalAliSullivanChatterjeeRistenpart19]

overview: PSI on asymmetric sets

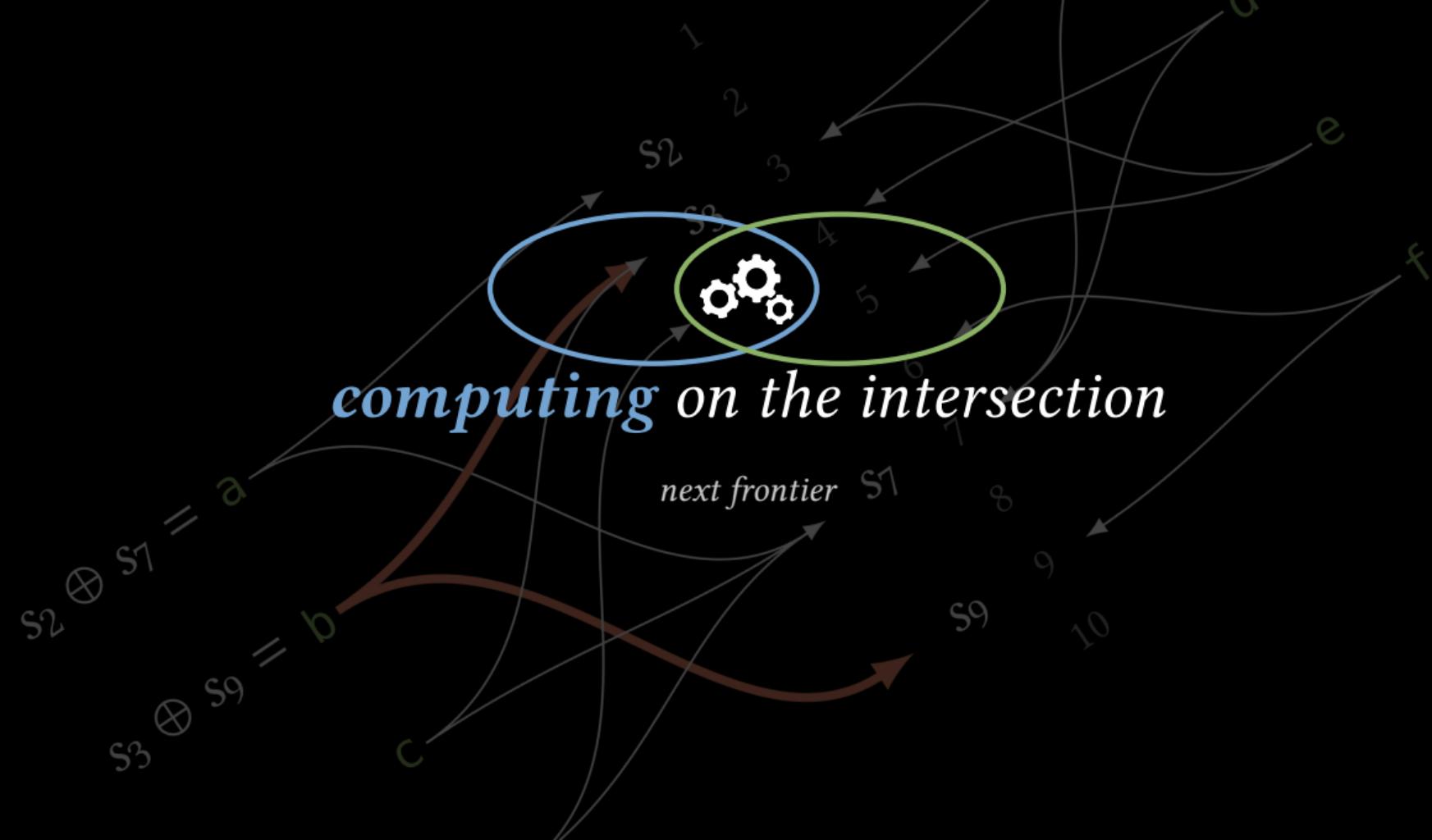
for 256 million vs 1000 items (no leakage):

offline setup: 33 seconds; 1 GB
discovery: 3 seconds; 6 MB

for 1 billion vs 100 items (under previous **leakage** scenario):

0.2 seconds; 1 MB

computing on the intersection

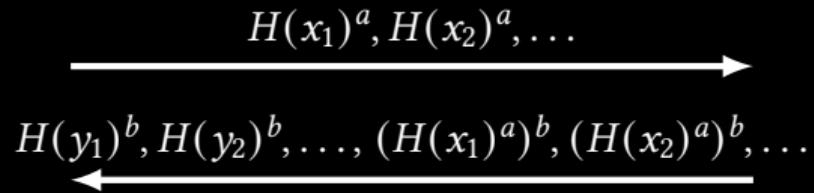


Alice

x_1, x_2, \dots

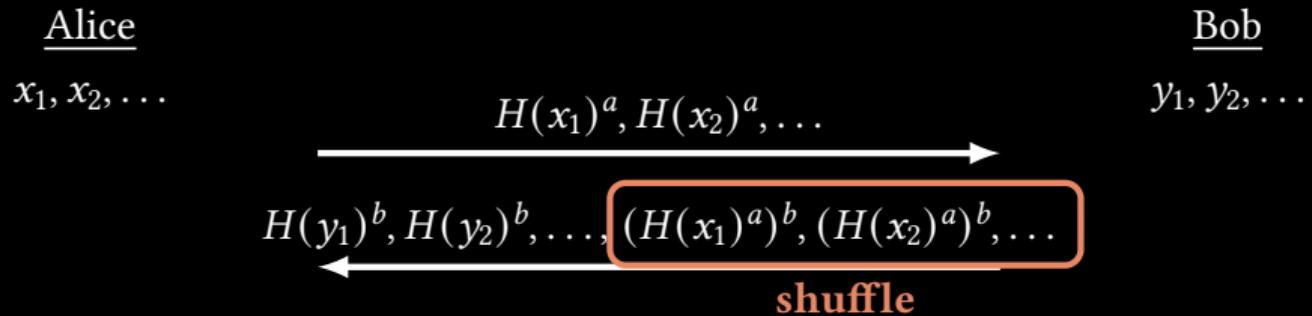
Bob

y_1, y_2, \dots



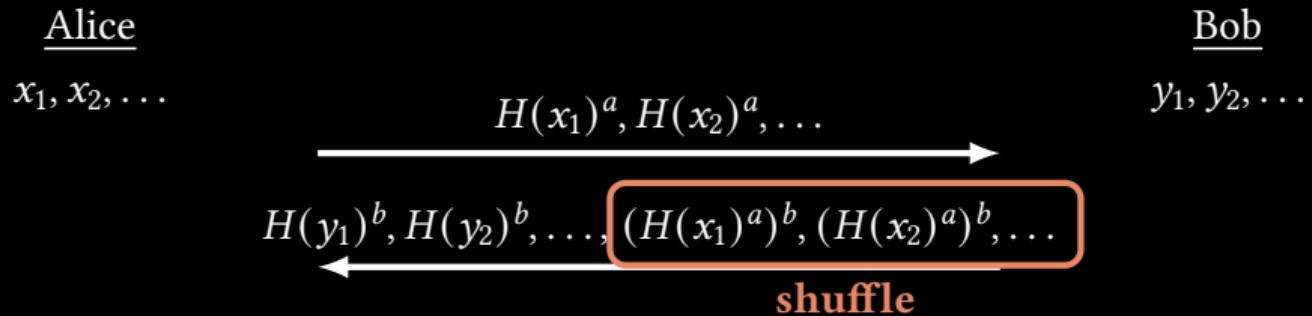
what is $X \cap Y$?

[HubermanFranklinHogg99]



what is $|X \cap Y|$?

[HubermanFranklinHogg99]



what is $|X \cap Y|$?

what about computing other functions of the intersection? what about large sets?

[HubermanFranklinHogg99]

state of the art

Alice

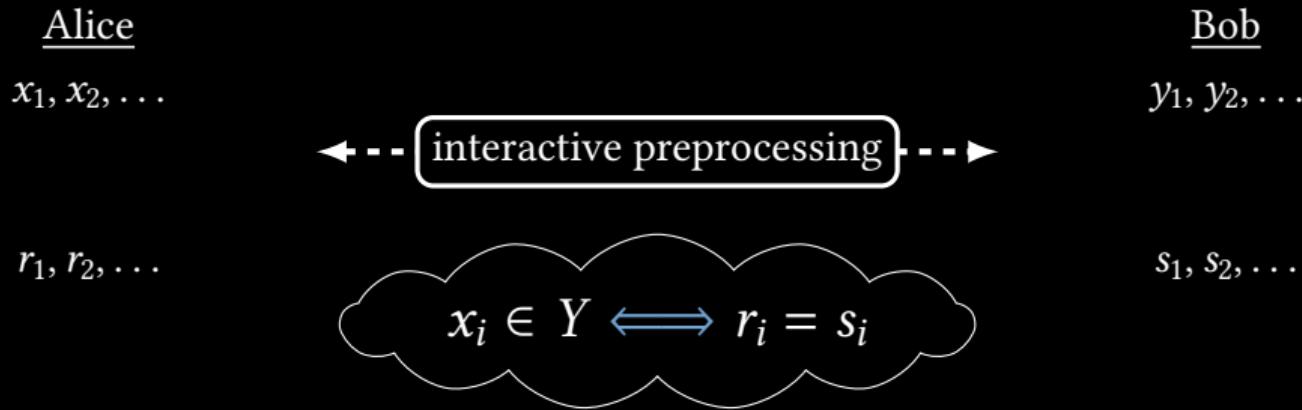
x_1, x_2, \dots

Bob

y_1, y_2, \dots

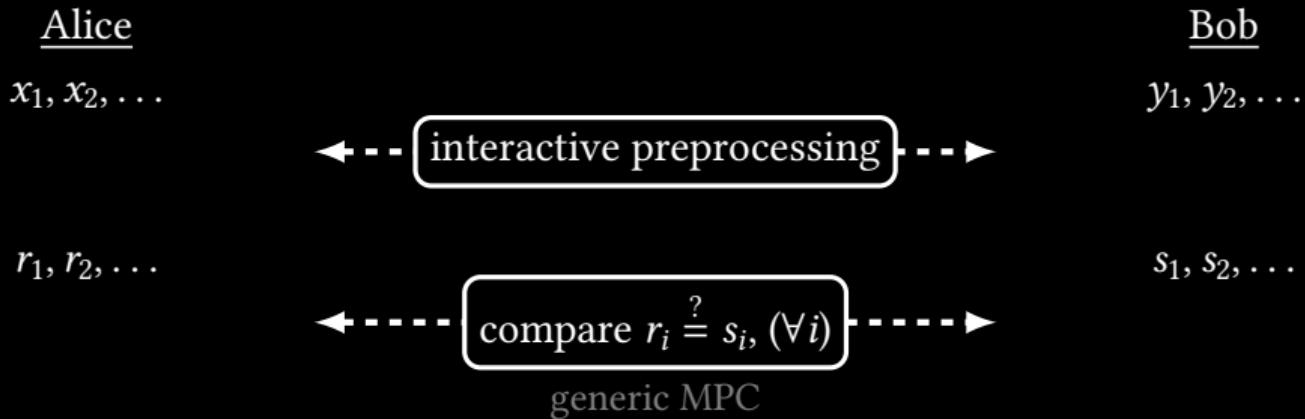
- ▶ Using $O(n)$ communication, reduce PSI to $O(n)$ comparisons (vs n^2)

state of the art



- ▶ Using $O(n)$ communication, reduce PSI to $O(n)$ comparisons (vs n^2)

state of the art



- ▶ Using $O(n)$ communication, reduce PSI to $O(n)$ comparisons (vs n^2)
- ▶ Perform the comparisons inside generic MPC \rightsquigarrow compute on the result

overview: computing on the intersection

for 1 million items:

2 minutes ; 2.5 GB

[PinkasSchneiderTkachenkoYanai19]

overview: computing on the intersection

for 1 million items:

2 minutes ; 2.5 GB

 30× plain PSI

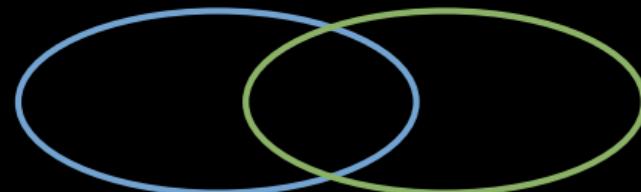
 20× plain PSI

[PinkasSchneiderTkachenkoYanai19]



PSI on **small sets** (hundreds)

- ▶ efficient! 0.1sec / 256 items
- ▶ based on Diffie-Hellman KA



PSI on **large sets** (millions)

- ▶ fast! 4sec / 1M items
- ▶ OT extension & hashing techniques



PSI on **asymmetric sets**

- ▶ huge challenges for practice
- ▶ allow leakage, preprocessing?



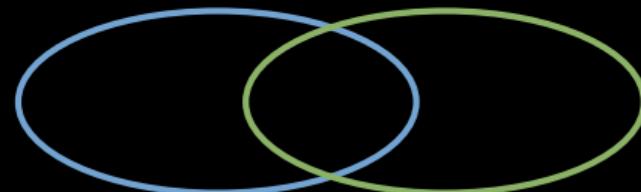
computing on the intersection

- ▶ many open problems
- ▶ 20-30× performance gap



PSI on **small sets** (hundreds)

- ▶ efficient! 0.1sec / 256 items
- ▶ based on Diffie-Hellman KA



PSI on **large sets** (millions)

- ▶ fast! 4sec / 1M items
- ▶ OT extension & hashing techniques

thank you!



PSI on **asymmetric sets**

- ▶ huge challenges for practice
- ▶ allow leakage, preprocessing?



computing on the intersection

- ▶ many open problems
- ▶ 20-30× performance gap