An Effective and Differentially Private Protocol for Secure Distributed Cardinality Estimation

(ACM Conference on Management of Data, June 2023)

Pinghui Wang[1] *, Chengjin Yang[1], Dongdong Xie[1], Junzhou Zhao[1], Hui Li[2], Jing Tao[1], Xiaohong Guan[1]

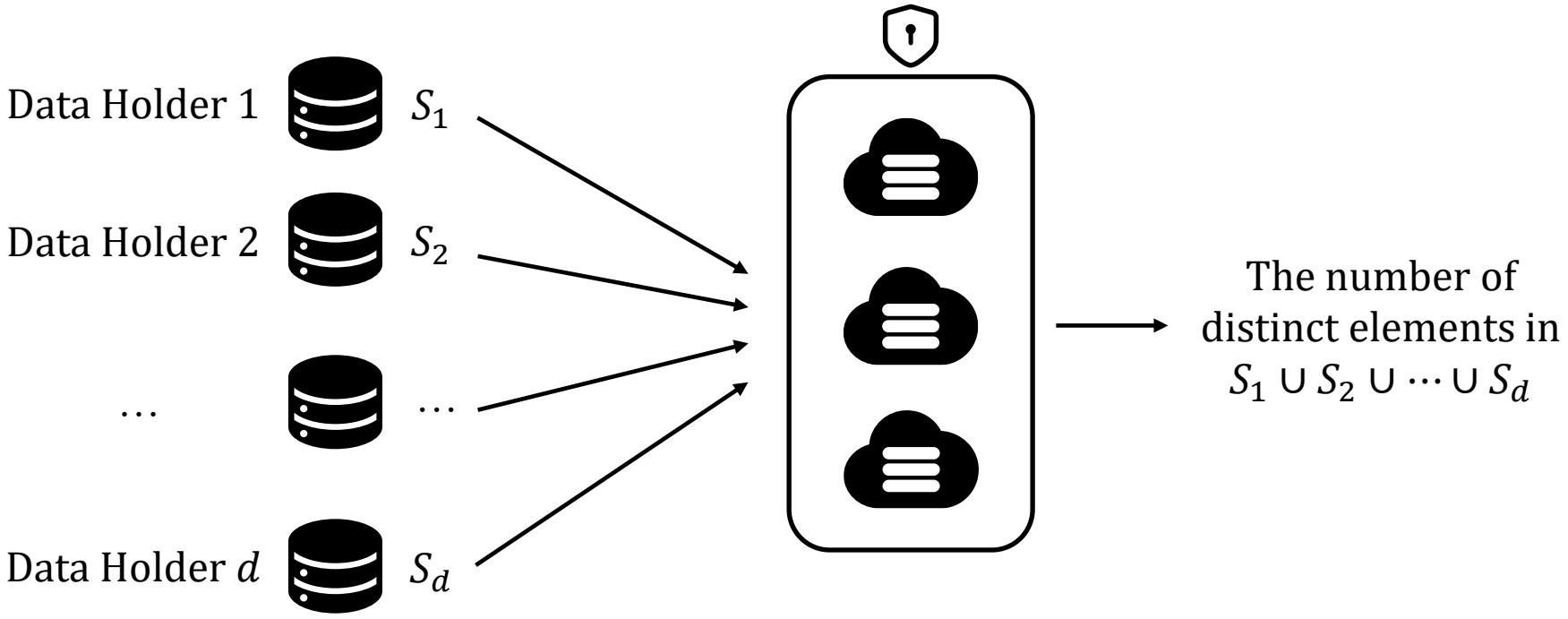
[1] Xi'an Jiaotong University
[2] Xidian University

Outline

- Introduction
- Preliminaries: FM sketch SPDZ Differential privacy
- Our protocol
- Conclusion

Introduction: the problem

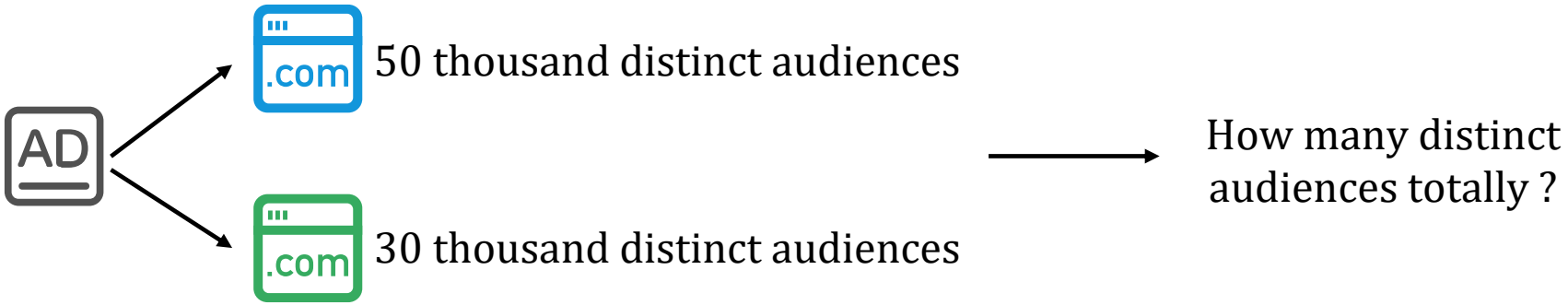
- Problem: counting the number of distinct elements in several data holders privately
- Private Distributed Cardinality Estimation (PDCE)



Introduction: applications

- Applications
  - Logistics Monitoring
  - Disease Epidemiology
  - Audience Reach Reporting
  - Collection of Internet Traffic Statistics

- Audience Reach Reporting



Introduction: our contributions

- Contribution 1
  - Reveal that the state-of-the-art protocol MPC-FM [1] is not differentially private, which is inconsistent with the claim in the original paper [1] .
- Contribution 2
  - Propose the new FMS sketch and a novel protocol *DP-DICE* for solving the PDCE problem.
  - Analyze the accuracy of FMS sketch and the security of DP-DICE.
- Contribution 3
  - Accelerates the computation speed on DHs by orders of magnitudes
  - Decreases estimation errors by several times to achieve the same security requirements.

[1] Changhui Hu, Jin Li, Zheli Liu, Xiaojie Guo, Yu Wei, Xuan Guang, Grigorios Loukides, and Changyu Dong. How to make private distributed cardinality estimation practical, and get differential privacy for free. In USENIX Security, 2021.

Introduction: problem formulation

- Cardinality: the number of distinct elements in a set

- Engagement parties:
  - d data holders (each holds a set S_i, e.g. website audiences)
  - c computational parties (help in counting, e.g. cloud servers)

- Goal:
  - Estimate n accurately (i.e. cardinality of $S_1 \cup S_2 \cup \cdots \cup S_d$)
  - Keep the existence of an element e in set S_i private (i.e. differential privacy)

Introduction: threat model

- Assumption 1.
  - Among the c computation parties (CPs), at most $c-1$ CPs is corrupted.
- Assumption 2.
  - The DHs will faithfully execute the protocol but may try to learn all possible information legally.
- Assumption 3.
  - The estimation result of cardinality n is publicly available to all the DHs, the CPs, and the adversary.

Outline

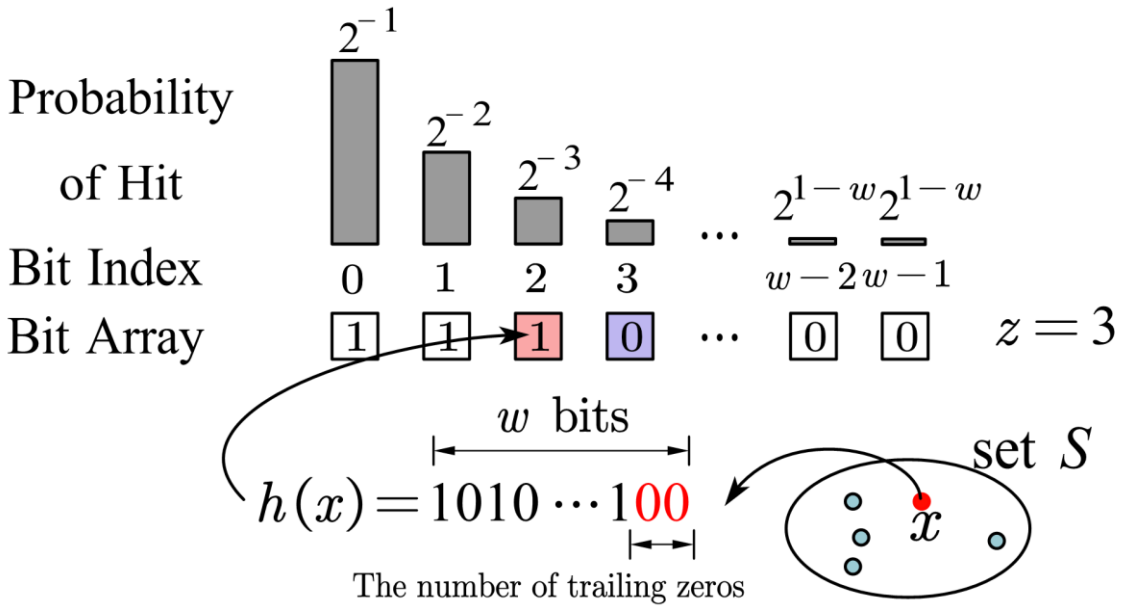
- Introduction
- Preliminaries: FM sketch SPDZ Differential privacy
- Our protocol
- Conclusion

Preliminaries: FM sketch

- Sketch: a compact data structure
- FM sketch: a classic data structure to efficiently estimate the cardinality (i.e., the number of distinct elements) of a large set S
- Memory space: $m \times w$
  - m bit arrays and each bit array has length w
  - (m usually $1024 - 8192$; $w = \log(n) + 4$, usually 32)

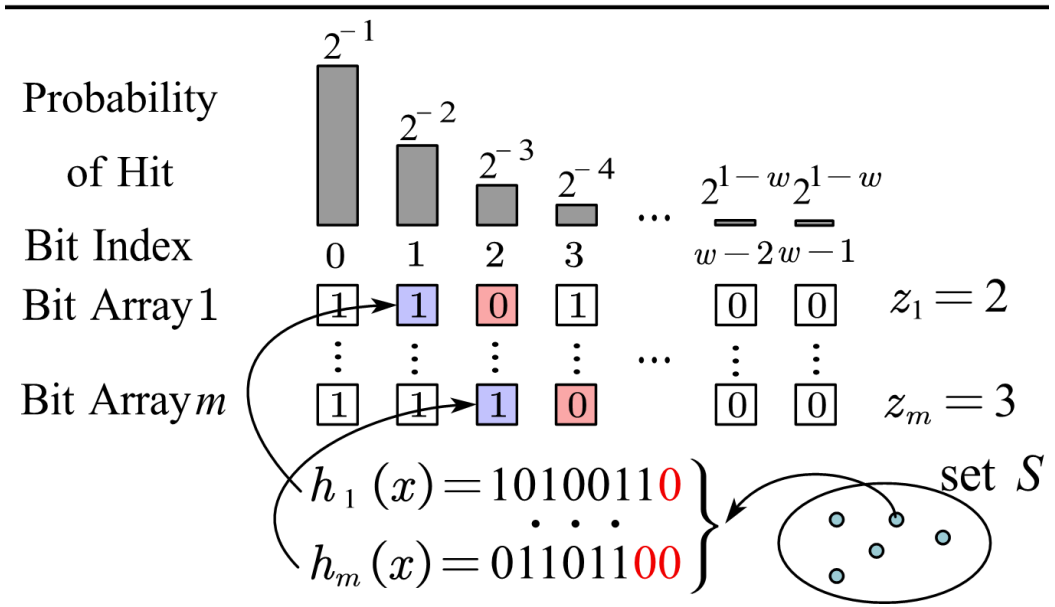
Preliminaries: FM sketch

- m bit arrays and each bit array has length w
- Take a single bit array (length w) as an example
- The index of first zero (i.e. z) reflects the cardinality



Preliminaries: FM sketch

- m bit arrays and each bit array has length w
- To further reduce estimation error, more bit arrays are used
- Each bit array has a different hash function



Estimator $\hat{n} = 2^{\sum_{i=1}^{m} z_i} / \varphi$

Preliminaries: FM sketch

- The standard estimation error: $\frac{\sqrt{\mathrm{Var}(\hat{n})}}{n} \approx \frac{1.0}{\sqrt{m}}$

- FM sketch can merge:
  - Knowing FM sketch F_1 of set S_1, FM sketch F_2 of set S_2
  - The FM sketch of set $S_1 \cup S_2$ is $F_1 \vee F_2$ ($\vee$ means bit-wise or operation)

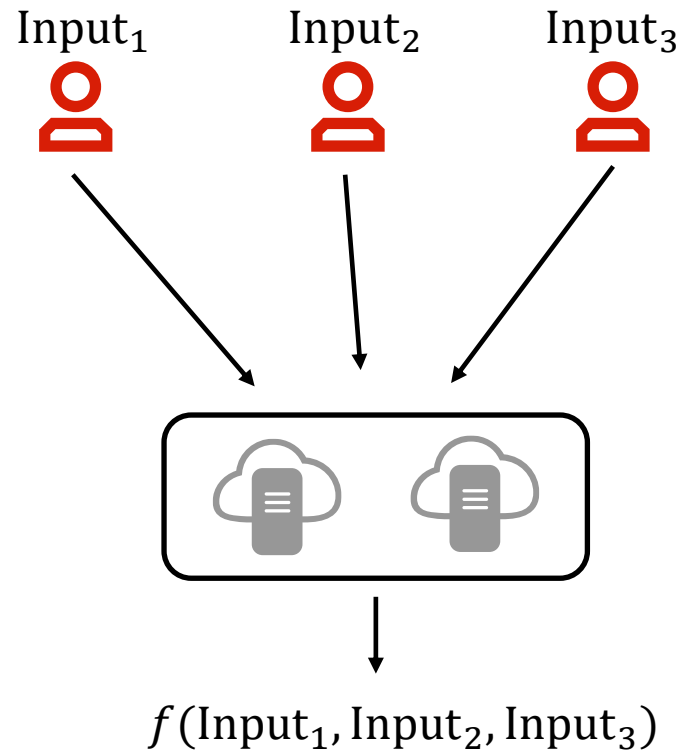
Preliminaries: SPDZ

A secure multi-party computation (MPC) protocol

- A group of parties each holding an input
- They compute a function together:

  $$f(\text{Input}_1, \cdots, \text{Input}_d)$$

  and only output the outcome
- They keep everyone's own input unknown to others



Preliminaries: Differential privacy

- Only outputting the outcome may also leak privacy

- e.g. privately counting #poor students in a class
  - Day 1: 10 students
  - Day 2: a new student join the class
  - Day 3: 11 students
  - Leaks the privacy of the new student!

Preliminaries: Differential privacy

- Differential privacy:
  - A randomized algorithm $f: D \to Y$ is ε-differential private if for all pairs of $x, x' \in D$ differ in at most one element, for all $E \subseteq Y$, we have:

$$e^{-\varepsilon} \leq \frac{P(f(x) \in E)}{P(f(x') \in E)} \leq e^{\varepsilon}$$

- A method to achieve differential privacy
  - Adding a noise before output
  - Take Laplace noise [3] as an example, the noise needed is:

$$\text{Lap}(\mu = 0, b = \frac{\Delta f}{\varepsilon})$$

where $\Delta f = \max | f(x) - f(x') |$ is called sensitivity

[3] Dwork C, McSherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis[C]//Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3. Springer Berlin Heidelberg, 2006: 265-284.

Outline

- Introduction
- Preliminaries: FM sketch SPDZ Differential privacy
- Our protocol
- Conclusion

Defects of FM sketch

- Updating a single element has time complexity O(m)
  - Each bit array needs a hash and an assignment
- Leaks privacy in distributed setting
  - Differentially private when only one data holder
  - Isn't differentially private in distributed settings [2]
- The noise needed for differential private is large
  - The sensitivity of $Z = \sum z$ is $m \times w$
  - The resulting estimation error is much larger than 100%

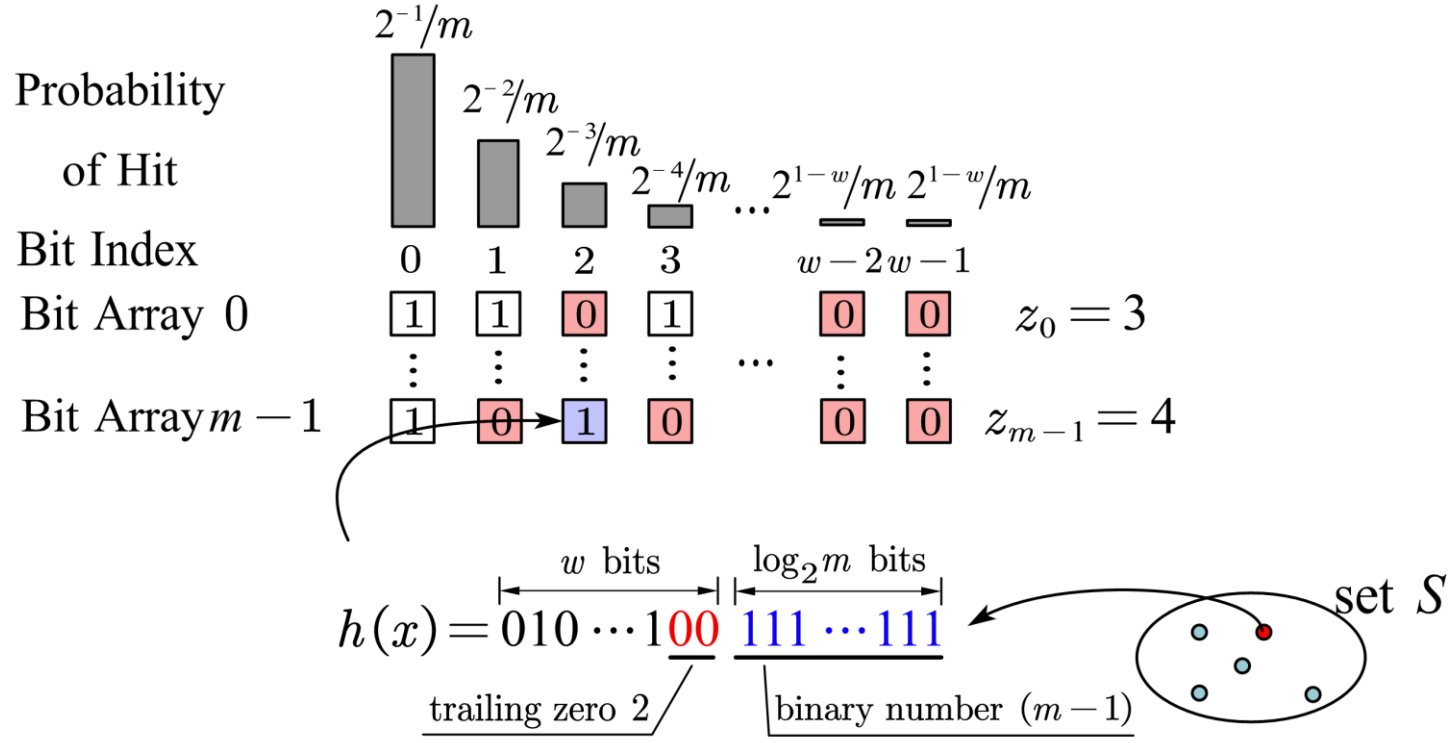
[2] Wang P, Yang C, Xie D, et al. An Effective and Differentially Private Protocol for Secure Distributed Cardinality Estimation[J]. arXiv preprint arXiv:2302.02158, 2023.

A new sketch: FMS sketch

- FMS sketch:
  - Using bucket hashing technique to accelerate
  - Using a new observation to estimate cardinality
- Orders of magnitudes quicker than FM sketch
- Small sensitivity, only 1
  - thus small noise
  - and small estimation error

A new sketch: FMS sketch

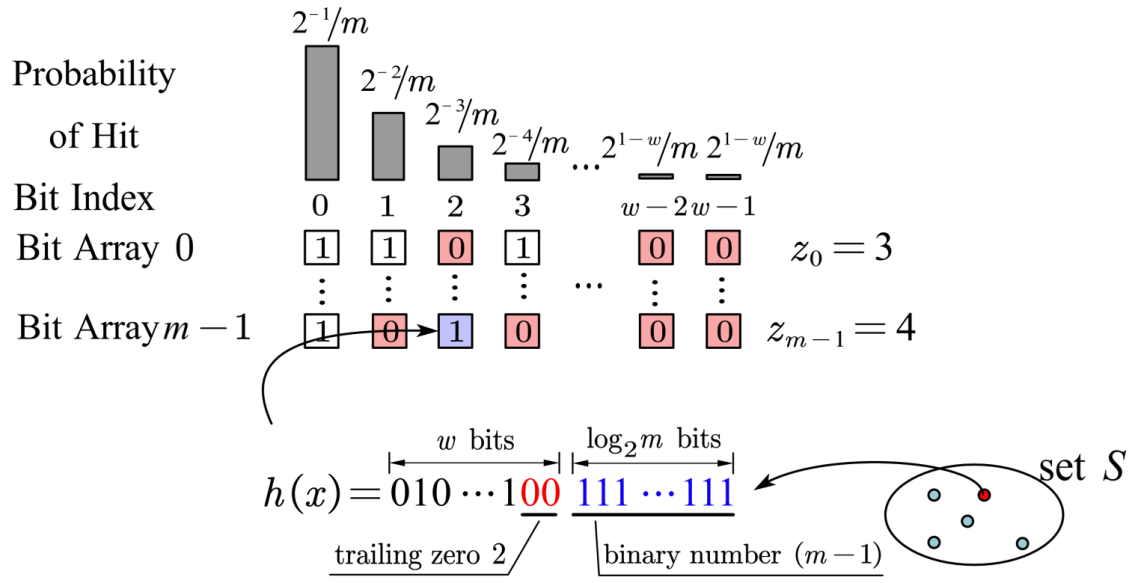
- FMS sketch:
  - Using bucket hashing technique
  - Using a new observation to estimate cardinality



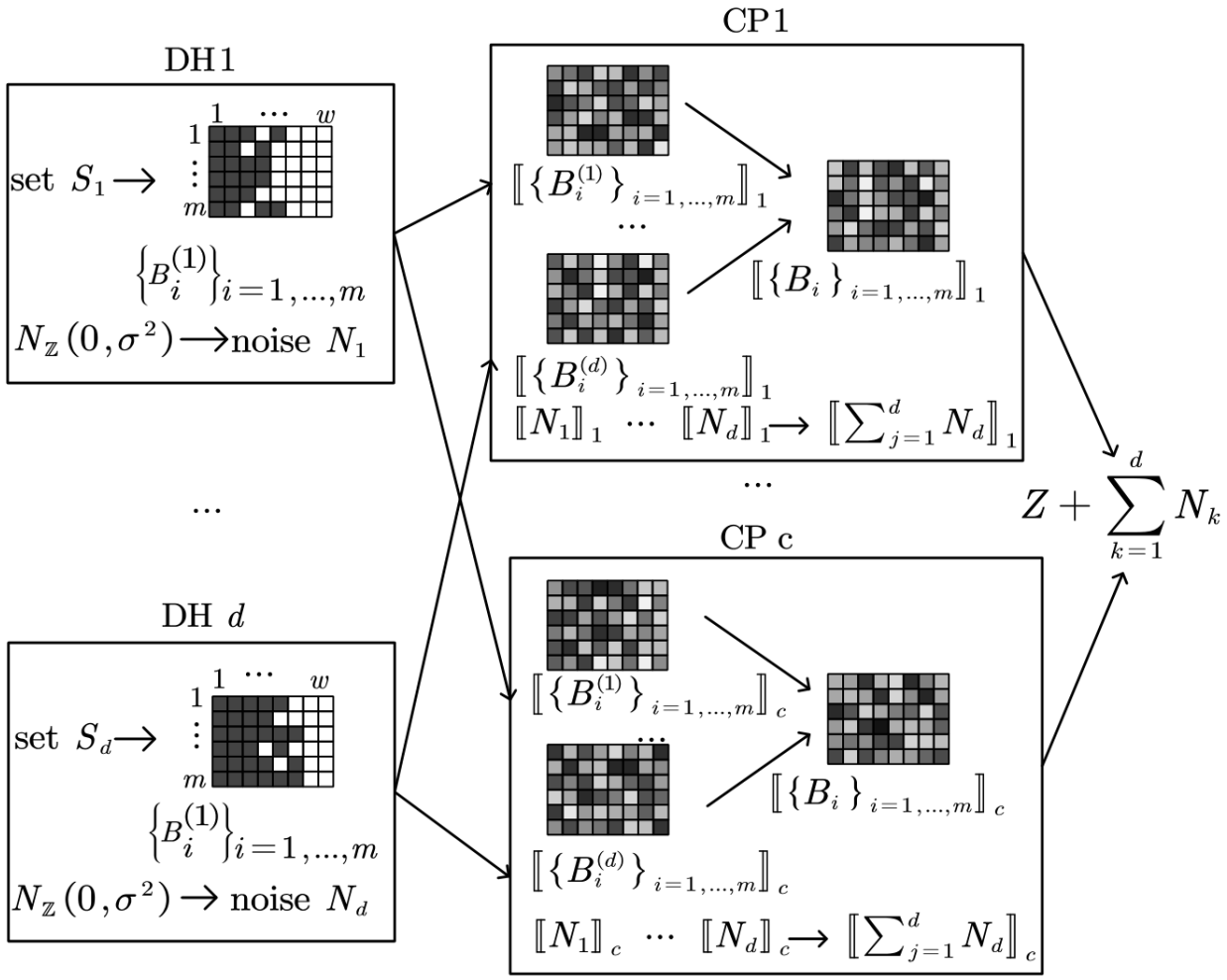
A new sketch: FMS sketch

Summary

- O(1) update procedure
- Sensitivity only 1
- Standard error: $\frac{\ln 2}{\sqrt{m}}$
- Easy to merge



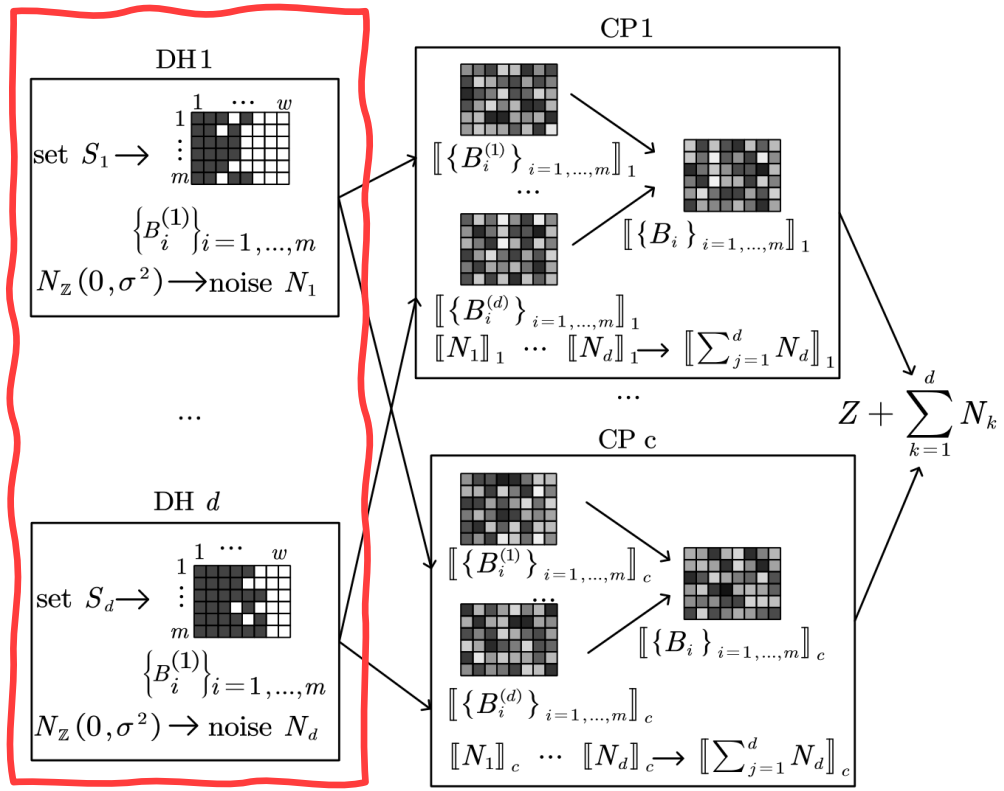
Our protocol: DP-DICE



- The novel FMS sketch
  - Estimate cardinality
- SPDZ
  - Ensure information security
- Noise
  - Ensure privacy

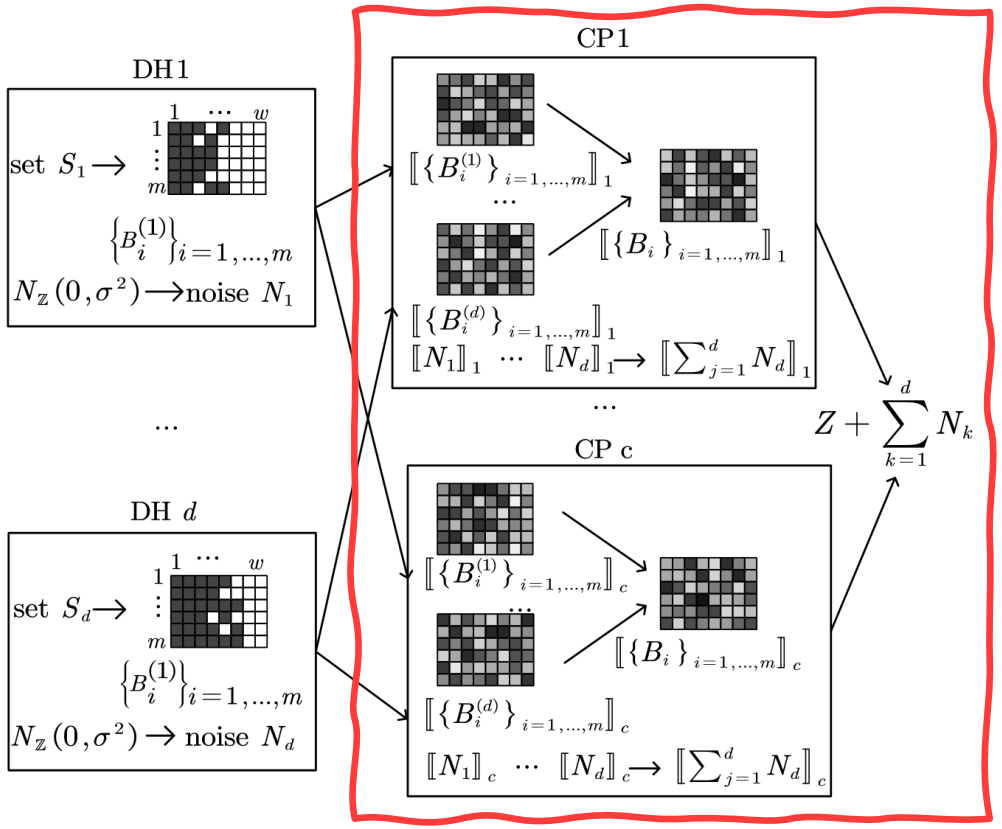
Our protocol: DP-DICE

- 1. Offline Preparation Phase
  - Parameter Initialization: $m, w, \mathbb{F}_p$ et.al
  - Random Number Generation: prepare for online phase
- 2. Data Collection Phase
  - FMS Sketch Generation
  - Noise Generation
  - Secure Data Sharing



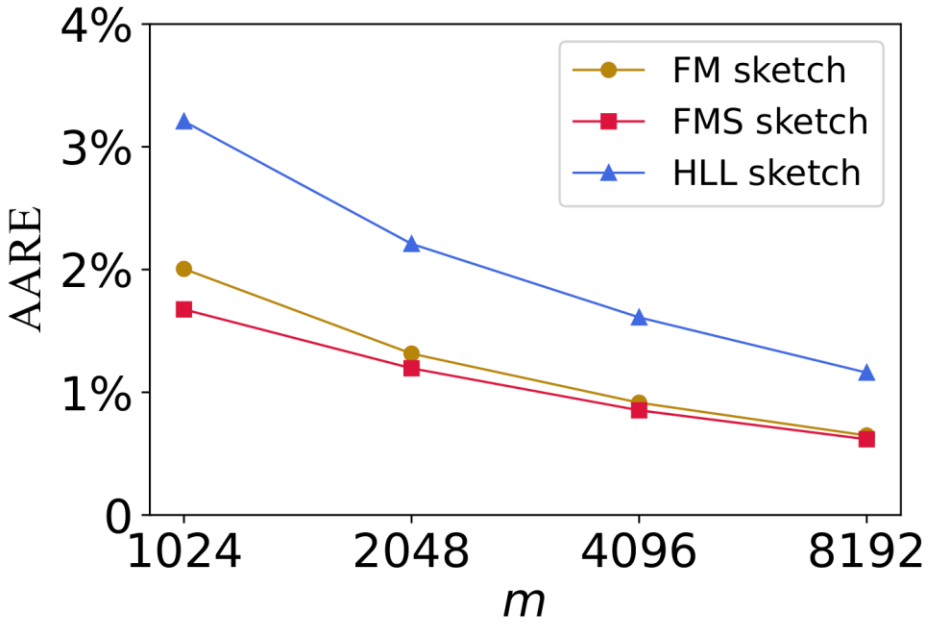
Our protocol: DP-DICE

- 3. Data Aggregation Phase
  - Merge FMS Sketches
  - Merge Noise Variables
  - Estimate Cardinality

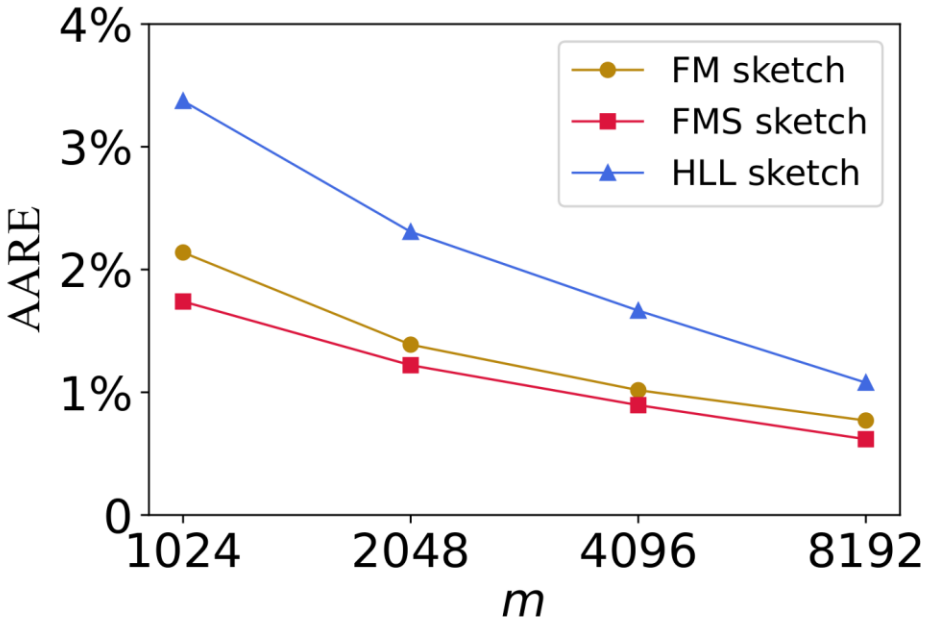


Experiments: FMS sketch is more accurate

- Accuracy of FM sketch, Hyperloglog sketch and FMS sketch when privacy is not concerned



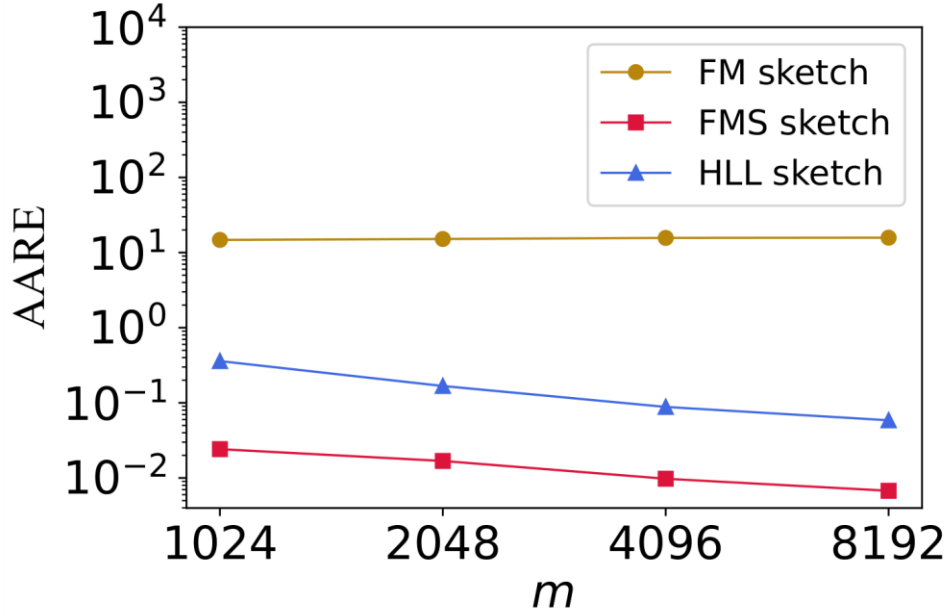
(c) AARE vs. m, where $n = 10^7$



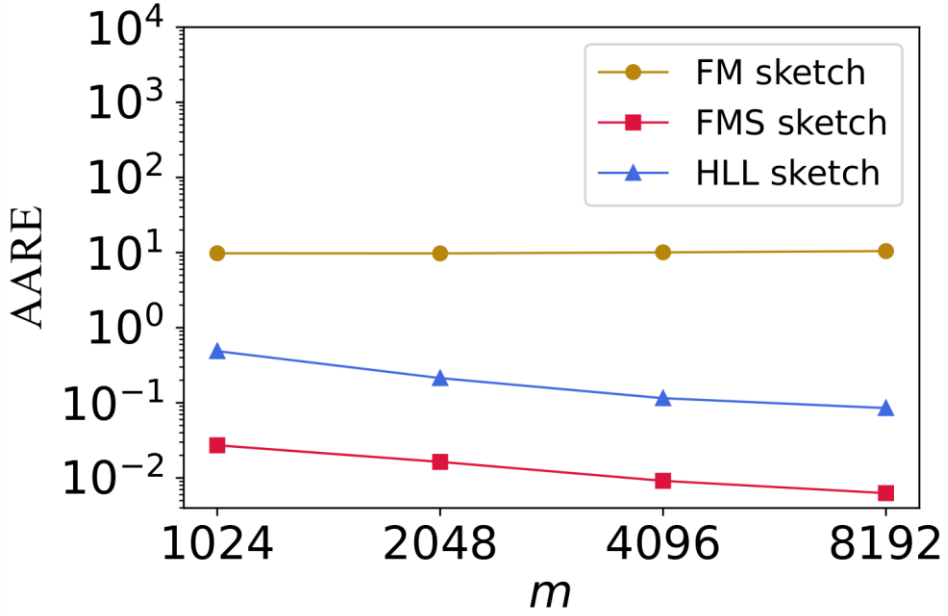
(d) AARE vs. m, where $n = 10^9$

Experiments: FMS sketch is more accurate

- Accuracy of FM sketch, Hyperloglog sketch and FMS sketch when privacy is concerned



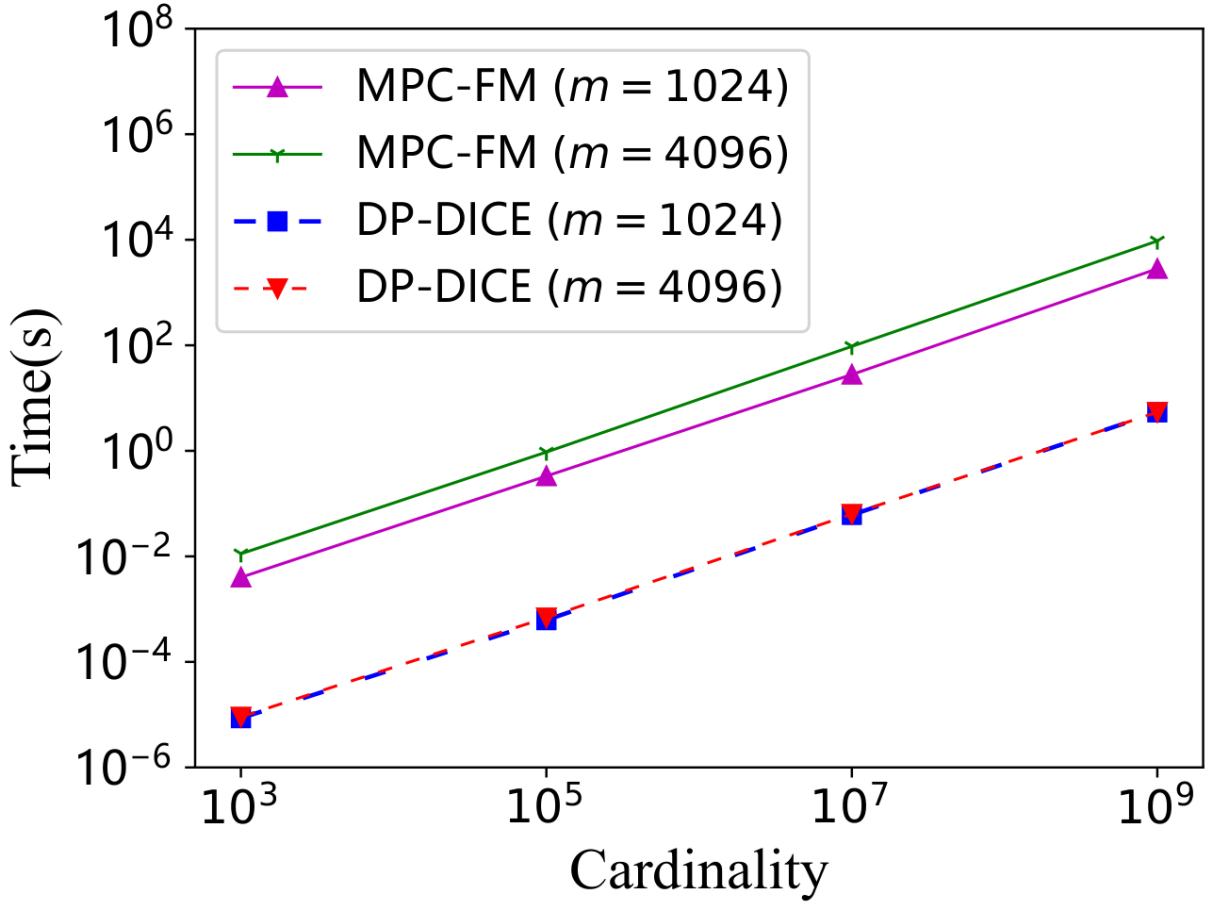
(g) AARE vs. m, where $n = 10^7$



(h) AARE vs. m, where $n = 10^9$

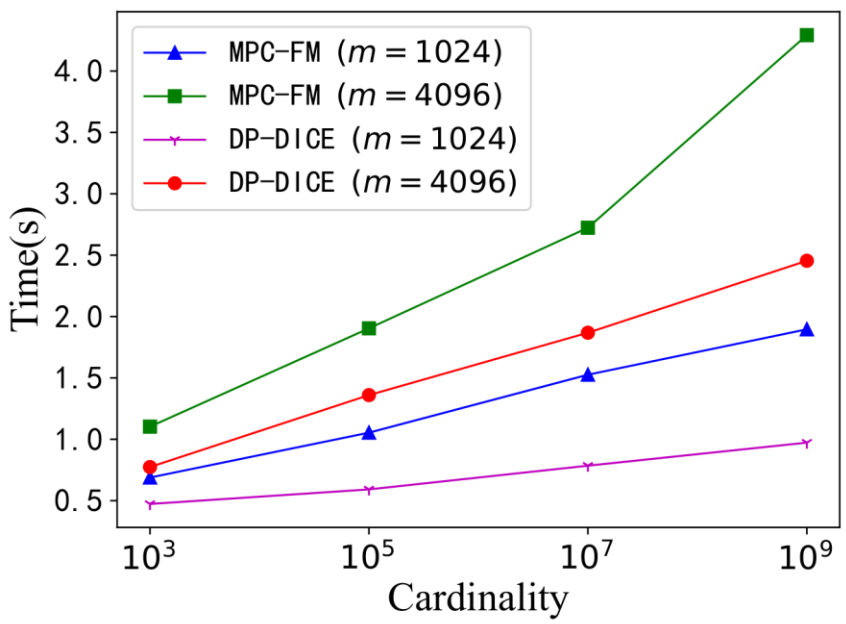
Experiments: DP-DICE is more efficient

- DHs' sketch generation time for our DP-DICE

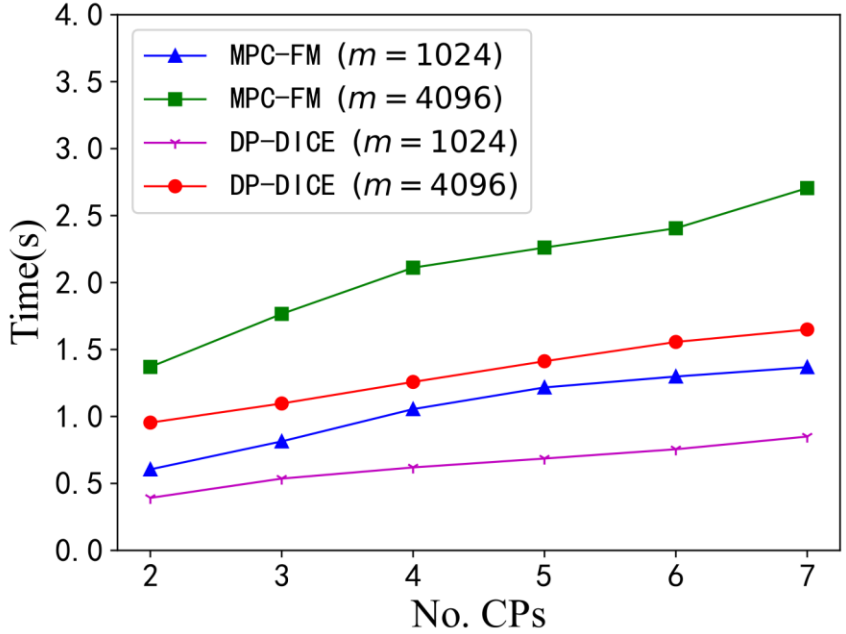


Experiments: DP-DICE is more efficient

- Online running time of our DP-DICE compared with MPC-FM (SOTA before ours)



(k) (WAN) online running time for different cardinalities



(l) (WAN) online running time for different numbers of CPs

Outline

- Introduction
- Preliminaries: FM sketch SPDZ Differential privacy
- Our protocol
- Conclusion

Conclusion

- Propose a novel sketch: FMS sketch
- Propose an efficient and secure protocol DP-DICE to solve the “Private distributed cardinality estimation” problem
- Advantages of DP-DICE
  - Is secure and differentially private
  - Reduces the estimation error by several times
  - Speeds up the sketch generation time by orders of magnitude

Thanks for attention!