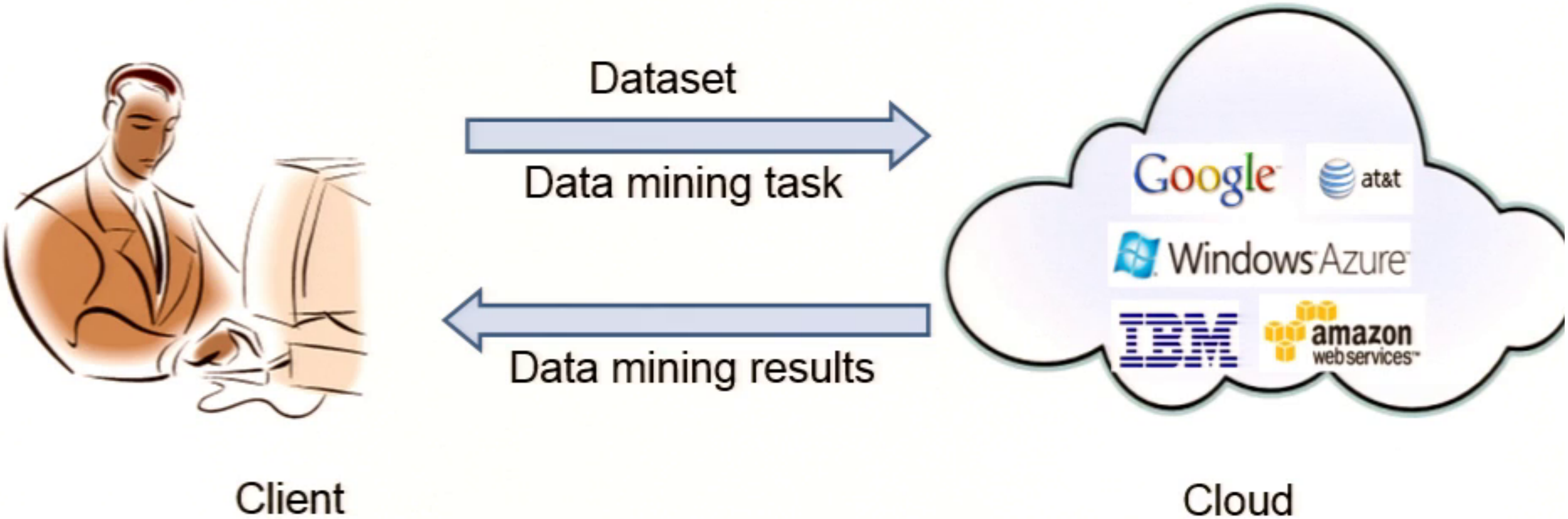
Result Integrity Verification of Outsourced Privacy-preserving Frequent Itemset Mining

Ruilin Liu, Wendy Hui Wang
Stevens Institute of Technology
Hoboken, NJ, USA

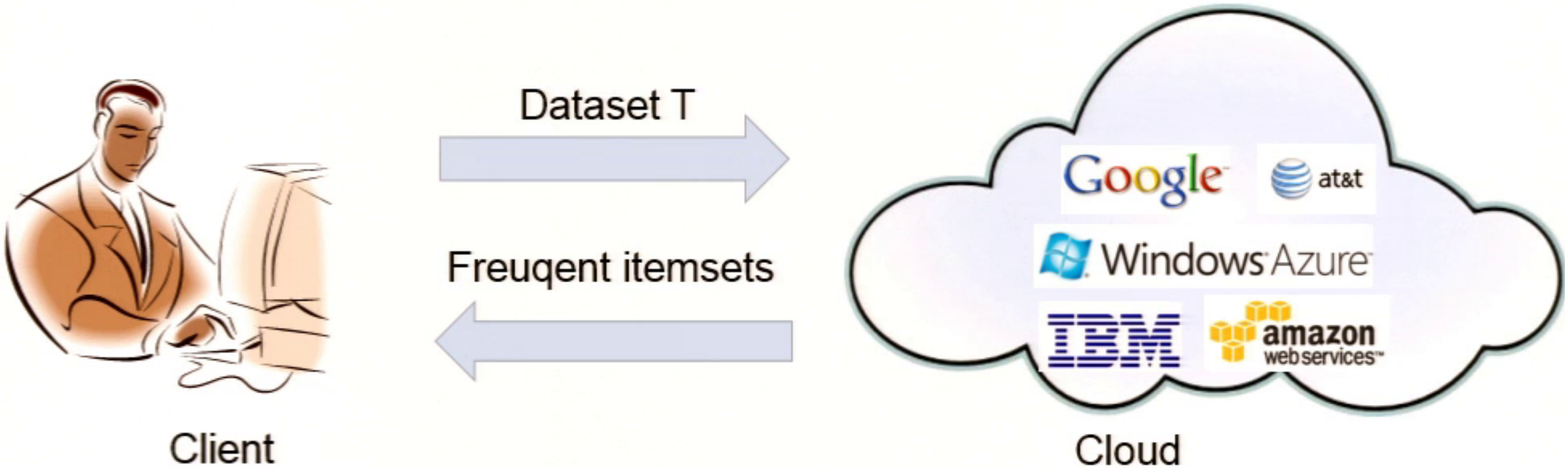
Supported by the National Science Foundation CAREER Grant #1350324

Data-Mining-as-A-Service (DMaS) Paradigm



We consider <u>frequent itemset mining</u> as the mining task

Security Issues of DMaS



Security concerns:

1. How to protect privacy of the data and the mining results?

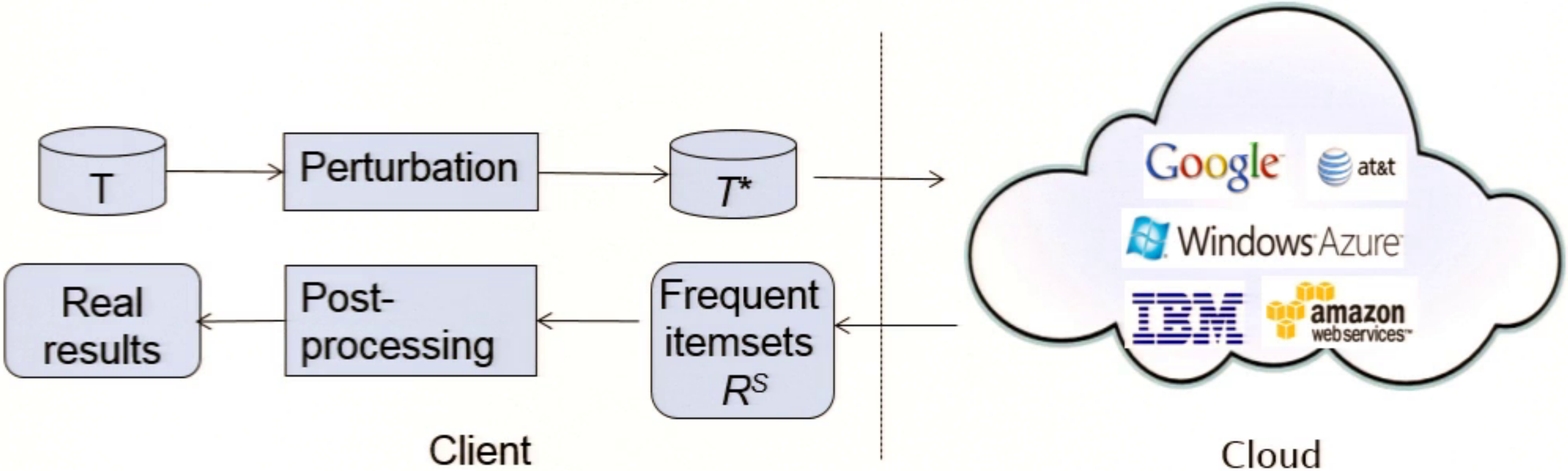
2. How to verify correctness/completeness of the mining results?

- Correctness: all returned itemsets are frequent
- Completeness: all frequent itemsets are returned

Existing Research

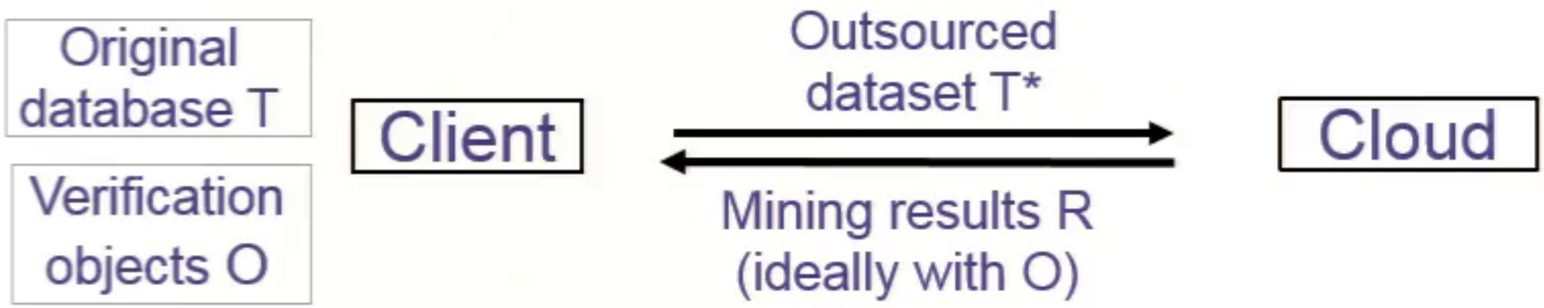
- Two parallel lines of research
  - Privacy-preserving mining (e.g., [1], [9])
  - Verification of outsourced data mining computations [2-7] (without any privacy protection)
- No work considered both privacy and result integrity verification in a unified framework

Privacy-Preserving Frequent Itemset Mining



- Select-a-size randomization approach [1]
  - Effect of randomization on itemset support:
    - The itemset support is a random variable following a given distribution
    - Frequent (infrequent resp.) itemsets may become infrequent (frequent, resp.)

Result Integrity Verification Methods



- Frequent itemset mining [2, 3] (without privacy protection):
  - **Verification preparation**
    - The client constructs artificial transactions Δ for verification objects
      - *Artificial frequent itemsets (FI)*: for completeness verification
      - *Artificial infrequent itemsets (II)*: for correctness verification
    - The client outsources T*= T + Δ.
  - **Verification**
    - The client verifies the completeness and correctness w.r.t. FI and II.

Verification Goal

- ***Correctness***: Precision $R_r = \frac{|R \cap R^S|}{|R^S|}$
- ***Completeness***: Recall $R_m = \frac{|R \cap R^S|}{|R|}$
  - R: frequent itemsets of T; R^S: mining results returned by the Cloud
- Verification goal
  - A verification method M can verify ***(α_1, β_1)-correctness*** if it has probability Pr ≥ α_1 to catch R^S whose precision $R_r \leq \beta_1$.
  - A verification method M can verify ***(α_2, β_2)-completeness*** if it has probability Pm ≥ α_2 to catch R^S whose recall $R_m \leq \beta_2$.
- Number of verification objects (FI and II) is decided by α_1, α_2, β_1, β_2.

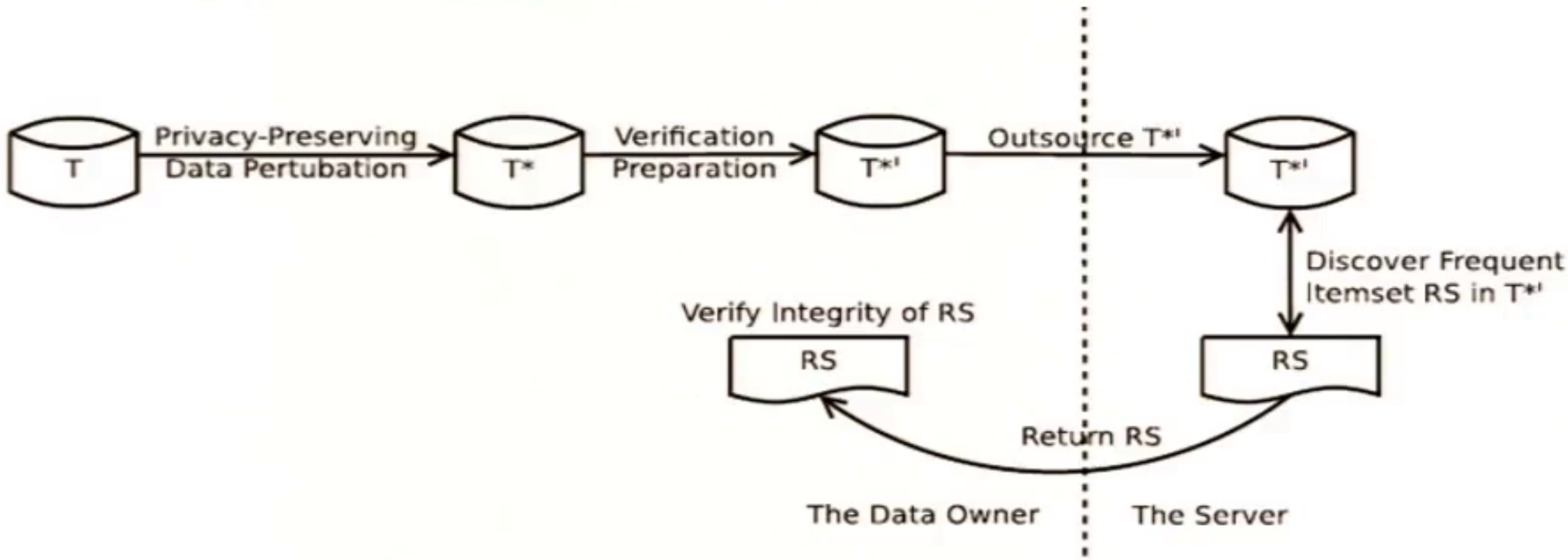
On the Marriage of Privacy and Result Integrity

- Two equal-important goals
  - Provable privacy guarantee
  - Robust result integrity guarantee (*(α_1, β_1)-correctness* and *(α_2, β_2)-completeness*)

- Challenges
  - Data-perturbation techniques lead to inaccurate mining results
  - It makes the Cloud's cheating behaviors harder to be caught.

Approach I

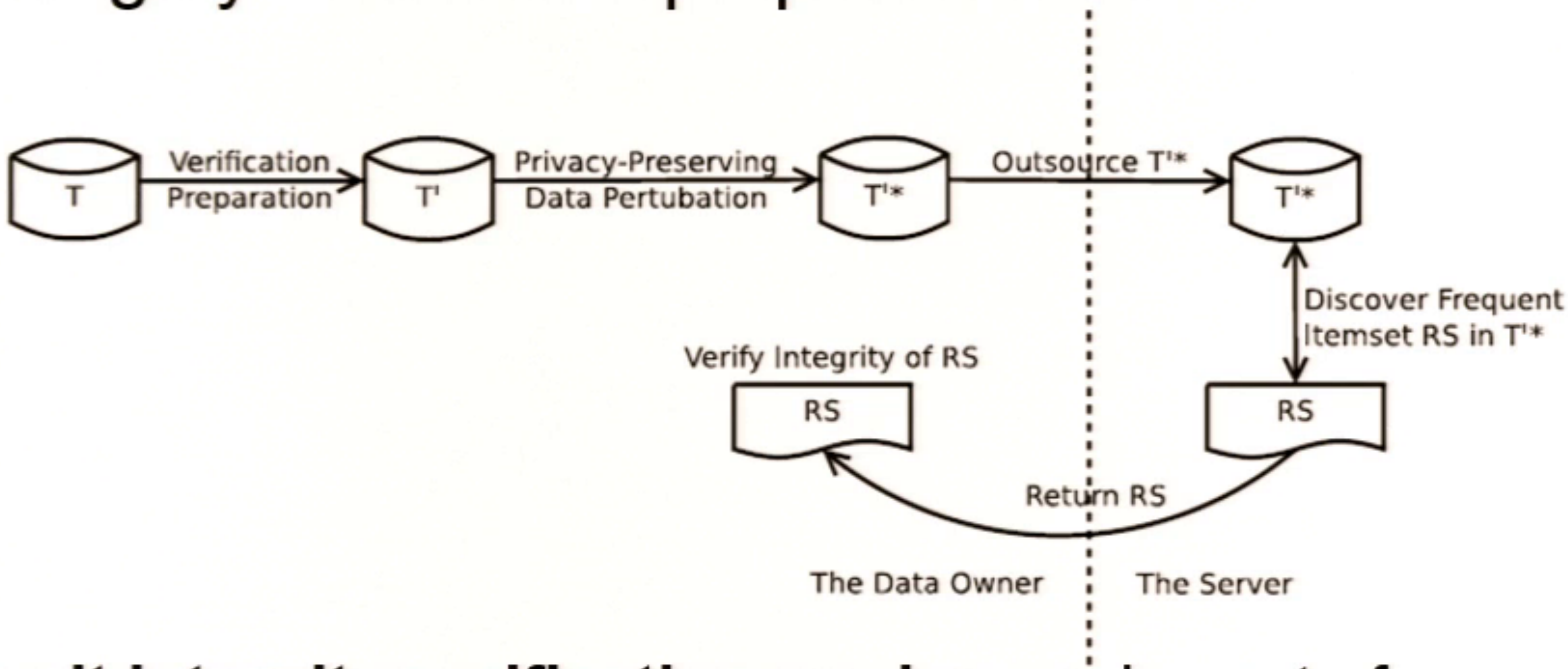
- Privacy-preserving data perturbation first



Privacy weakness: inserting artificial transactions constructed without any respect to privacy may lead to new privacy vulnerabilities.

Approach II

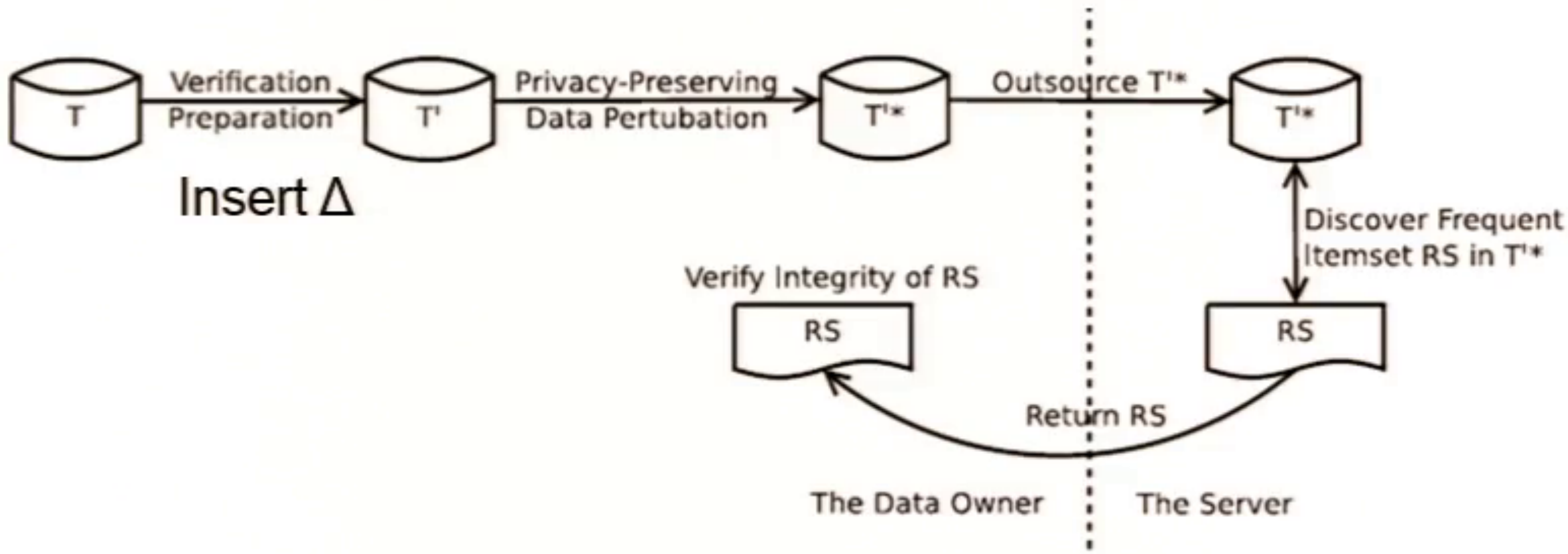
- Integrity verification preparation first



Result integrity verification weakness: impact of perturbation on verification objects

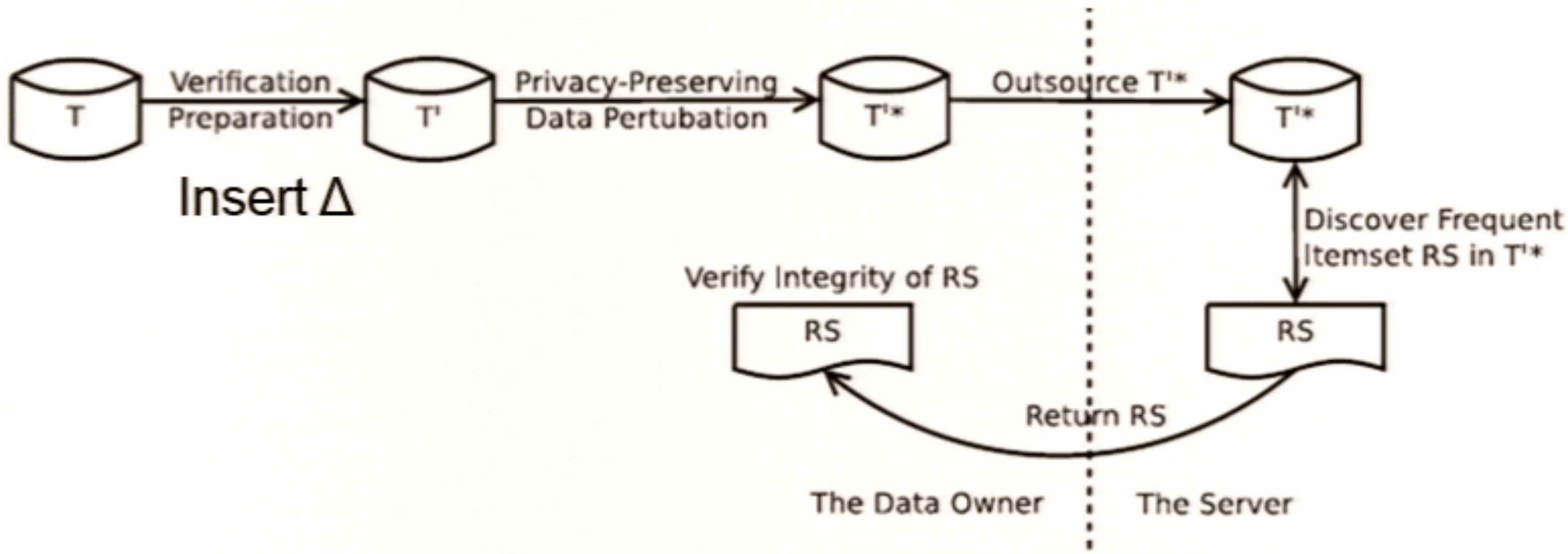
- Artificial frequent itemsets (FI) can turn to be infrequent.
- Artificial infrequent itemset (II) can turn to be frequent.

A Deeper Look of Approach II



- **Verification Preparation**: construction artificial transactions Δ that contains v_1 number of FI and v_2 number of II.
- **Privacy protection**: Apply *Select-A-Size* data pertubation [1].
- **Verification**: check if R^S contains at least r_1 FI and at most r_2 II.

A Deeper Look of Approach II



- **Verification Preparation**: construction artificial transactions Δ that contains v_1 number of FI and v_2 number of II.
- **Privacy protection**: Apply *Select-A-Size* data pertubation [1].
- **Verification**: check if R^S contains at least r_1 FI and at most r_2 II.

Challenges:

1. How to construct FI and II?
2. What is the appropriate value of v_1, v_2, r_1 and r_2 for *(α_1, β_1)-correctness and (α_2, β_2)-completeness*?

Our Contributions

- Design of efficient algorithms to construct verification objects (FI and II)
- Formal analysis of the probabilistic integrity guarantee
- Formal analysis of privacy guarantee

Probability Reasoning of Change of (In)Frequentness of FIs/IIs

Case	Itemset constructed by verification preparation	Itemset after data perturbation	Itemset in R^S	Reason	Probability
1	Frequent	Frequent	Y	True Positive	$Pr(FI \rightarrow F) * \beta_1$
2	Frequent	Frequent	N	Cheat on completeness	$Pr(FI \rightarrow F) * (1-\beta_2)$
3	Frequent	Infrequent	Y	Cheat on correctness	$Pr(FI \rightarrow I) * (1-\beta_1)$
4	Frequent	Infrequent	N	False Negative (by pertubation)	$Pr(FI \rightarrow I) * \beta_2$
5	Infrequent	Frequent	Y	False Positive (by pertubation)	$Pr(II \rightarrow F) * \beta_1$
6	Infrequent	Frequent	N	Cheat on completeness	$Pr(II \rightarrow F) * (1-\beta_2)$
7	Infrequent	Infrequent	Y	Cheat on correctness	$Pr(II \rightarrow I) * (1-\beta_1)$
8	Infrequent	Infrequent	N	True negative	$Pr(II \rightarrow I) * \beta_2$

α_1, β_1: for (α_1, β_1)-correctness; α_2, β_2: for (α_2, β_2)-completeness

Probability Reasoning of Change of (In)Frequentness of FIs/IIs

Case	Itemset constructed by verification preparation	Itemset after data perturbation	Itemset in R^S	Reason	Probability
1	Frequent	Frequent	Y	True Positive	$Pr(FI \rightarrow F) * \beta_1$
2	Frequent	Frequent	N	Cheat on completeness	$Pr(FI \rightarrow F) * (1-\beta_2)$
3	Frequent	Infrequent	Y	Cheat on correctness	$Pr(FI \rightarrow I) * (1-\beta_1)$
4	Frequent	Infrequent	N	False Negative (by pertubation)	$Pr(FI \rightarrow I) * \beta_2$
5	Infrequent	Frequent	Y	False Positive (by pertubation)	$Pr(II \rightarrow F) * \beta_1$
6	Infrequent	Frequent	N	Cheat on completeness	$Pr(II \rightarrow F) * (1-\beta_2)$
7	Infrequent	Infrequent	Y	Cheat on correctness	$Pr(II \rightarrow I) * (1-\beta_1)$
8	Infrequent	Infrequent	N	True negative	$Pr(II \rightarrow I) * \beta_2$

α_1, β_1: for (α_1, β_1)-correctness; α_2, β_2: for (α_2, β_2)-completeness

Probability Reasoning of Change of (In)Frequentness of FIs/IIs

Case	Itemset constructed by verification preparation	Itemset after data perturbation	Itemset in R^S	Reason	Probability
1	Frequent	Frequent	Y	True Positive	$Pr(FI \rightarrow F) * \beta_1$
2	Frequent	Frequent	N	Cheat on completeness	$Pr(FI \rightarrow F) * (1-\beta_2)$
3	Frequent	Infrequent	Y	Cheat on correctness	$Pr(FI \rightarrow I) * (1-\beta_1)$
4	Frequent	Infrequent	N	False Negative (by pertubation)	$Pr(FI \rightarrow I) * \beta_2$
5	Infrequent	Frequent	Y	False Positive (by pertubation)	$Pr(II \rightarrow F) * \beta_1$
6	Infrequent	Frequent	N	Cheat on completeness	$Pr(II \rightarrow F) * (1-\beta_2)$
7	Infrequent	Infrequent	Y	Cheat on correctness	$Pr(II \rightarrow I) * (1-\beta_1)$
8	Infrequent	Infrequent	N	True negative	$Pr(II \rightarrow I) * \beta_2$

α_1, β_1: for (α_1, β_1)-correctness; α_2, β_2: for (α_2, β_2)-completeness

Pr(*FI*→F) and Pr(*FI*→I)

- FI remains frequent after perturbation (case 1 & 2):

$$Pr(FI \to F) = \sum_{i=min_{sup}}^{N} Pr[supp_{T'*}(FI) = i],$$

- FI turns to be infrequent after perturbation (case 3 & 4) :

$$Pr(FI \to I) = \sum_{i=0}^{min_{sup}-1} Pr[supp_{T'*}(FI) = i],$$

where

$$Pr[supp_{T'*}(FI) = k] = \sum_{j=0}^{\min(k,a)} \binom{a}{j} (p_{\ell}^{m}[\ell \to \ell])^{j} \times \ell \to \ell])^{a-j} \times \binom{N}{k-j} (\rho_{m}^{\ell})^{k-j} \times (1 - \rho_{m}^{\ell})^{N-k+j}.$$

a: number of artificial transactions.

Pr(*II*→F) and Pr(*II*→I)

- II turns to be frequent after perturbation (case 5 & 6):

$$Pr(II \rightarrow F) = \sum_{i=min_{sup}}^{N} Pr[supp_{T'*}(II) = i],$$

- II remains infrequent after perturbation (case 7 & 8):

$$Pr(II \rightarrow I) = \sum_{i=0}^{min_{sup}-1} Pr[supp_{T'*}(II) = i],$$

Where:

$$Pr[supp_{T'*}(II) = k] \binom{N}{k} (\rho_m^{\ell})^k (1 - \rho_m^{\ell})^{N-k}.$$

Number of FI and II for Verification Preparation

- v_1: # of FI
- v_2: # of II
- The number of FI and II

$$\nu_1 = \log_{[(1-Pr[FI \rightarrow F])\beta_2]}(1-\alpha_2) + \log_{[(Pr[FI \rightarrow F])\beta_1]}(1-\alpha_1)$$

$$\nu_2 = \log_{[(1-Pr[II \rightarrow F])\beta_2]}(1-\alpha_2) + \log_{[(Pr[II \rightarrow F])\beta_1]}(1-\alpha_1)$$

Number of FI and II for Verification

- r1: expected # of FI in the returned result R^S
- r2: expected # of II in the returned result R^S
- r_1 and r_2 are computed as:

$$r_1 = \log_{(\beta_1 \times (Pr[FI \rightarrow F])}(1 - \alpha_1)$$

$$r_2 = \log_{(\beta_2 \times (Pr[II \rightarrow F])}(1 - \alpha_2)$$

Post-Processing

- Post-processing by the client
  - Remove FI and II
  - Recover real supports of real frequent itemsets

Complexity Analysis

- Client side
  - Preparation: $O(|FI|+|II|)$
  - Verification: $O(|FI|+|II|)$
  - Post-processing: $O(|R^S|)$
- Cloud side
  - $O(2^{|I|+|I1|+|I2|})$
    - I: number of unique items in T;
    - I_1/I_2: number of unique items in FI/II.

Privacy Analysis

- Our method is ε-*private*
  - For any transaction $t \in T$, and any itemset $A \subseteq t^*$, where t^* is constructed from t after perturbation

$$Pr\,[a \in t \mid A \subseteq t^*] < \varepsilon,$$

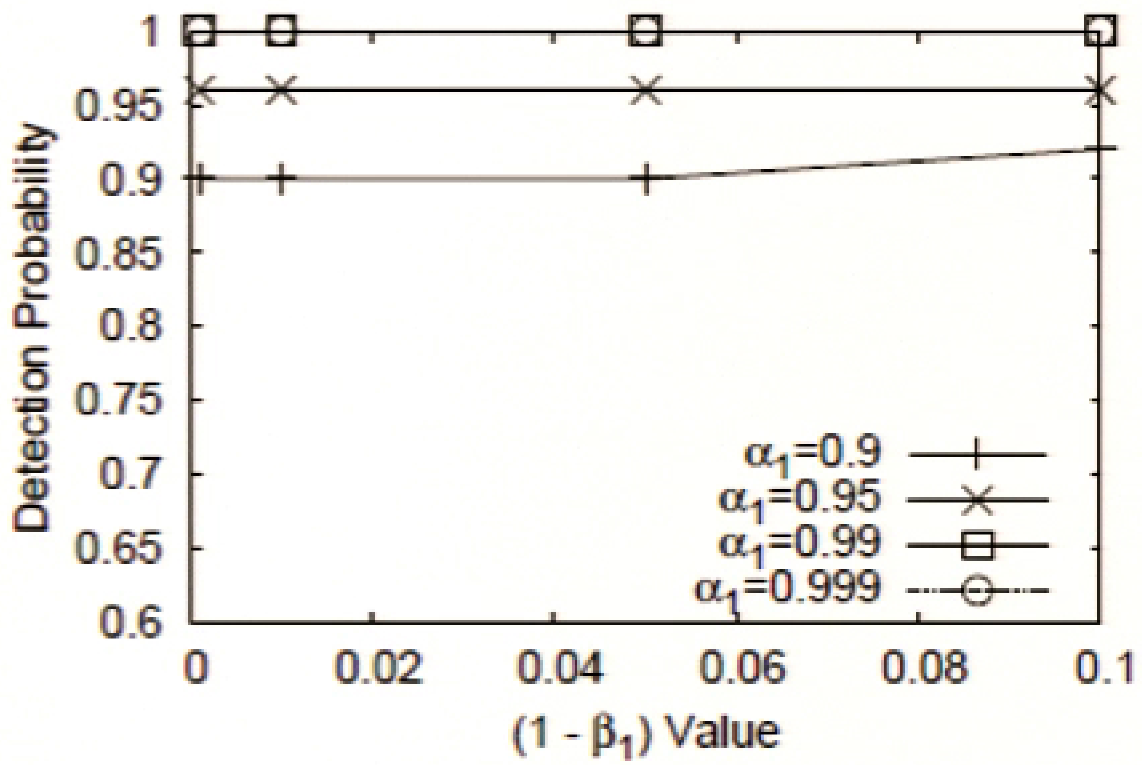
for any item $a \in t$.

Experiments

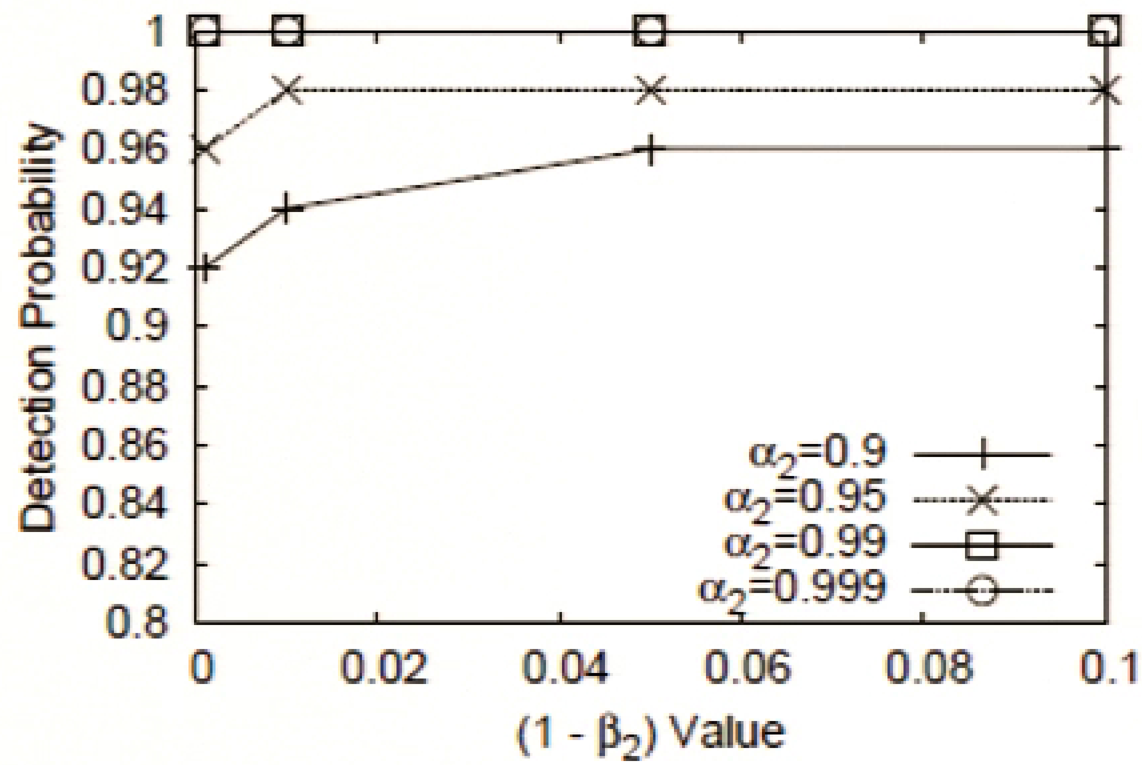
Datasets

Dataset	NASA-HTTP	Retail
# of transactions	39531	88162
# uf unique items	22458	16470
max length of transactions	112	74
min length of transactions	1	1
min_{sup}	1000	10
# of frequent itemsets	4156264	189400

Detection Probability

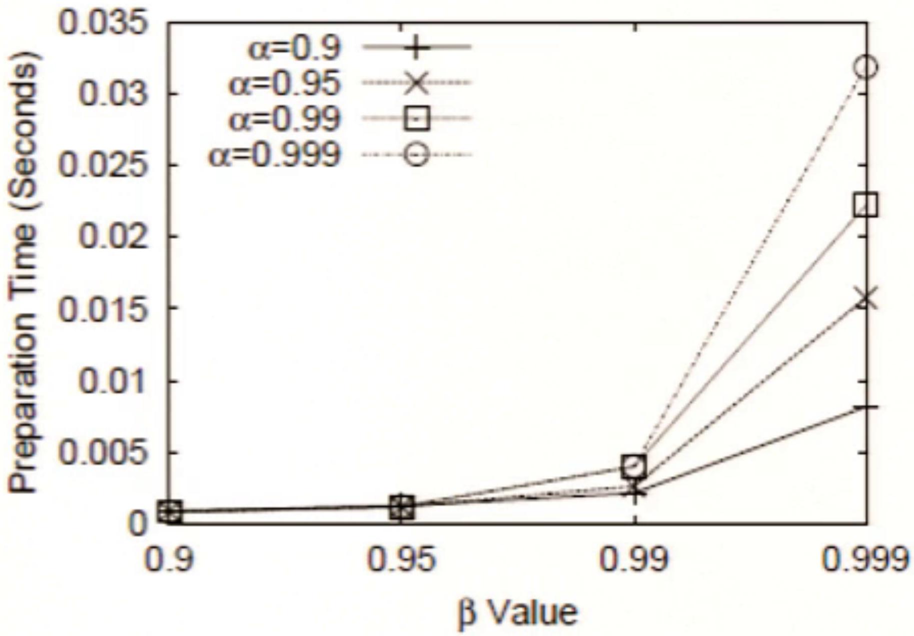


(a) Correctness Verification

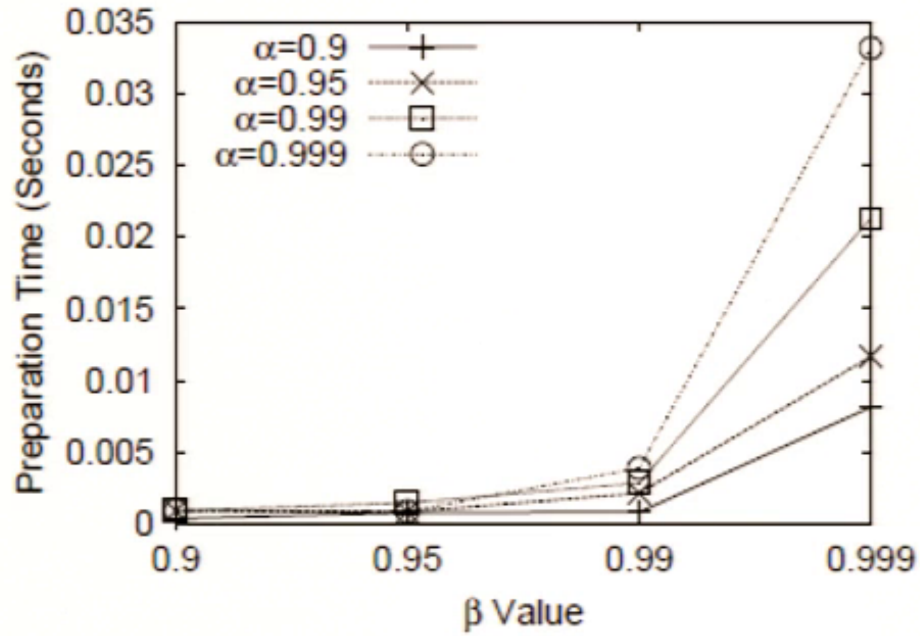


(b) Completeness Verification

Verification Preparation Time

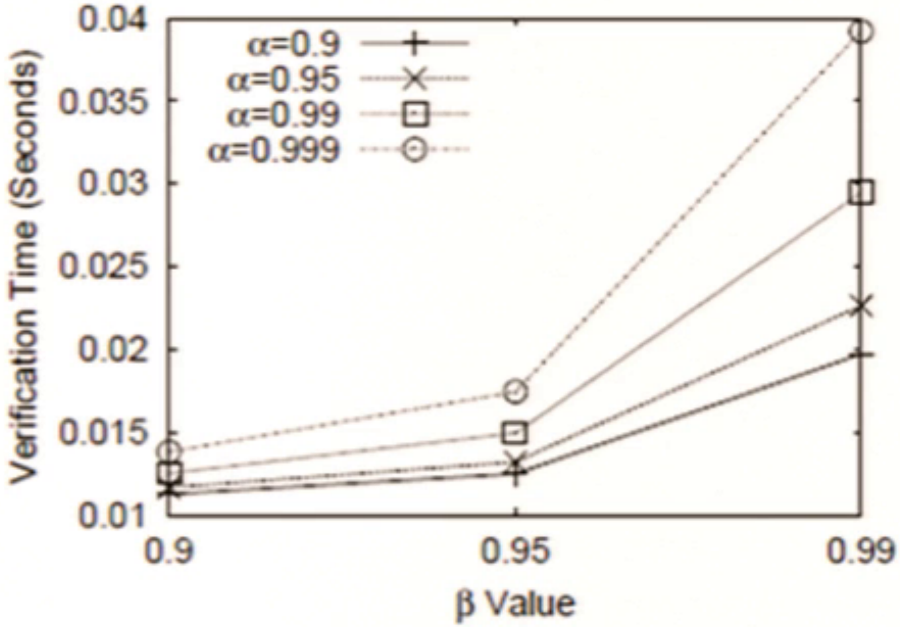


(a) Retail Dataset

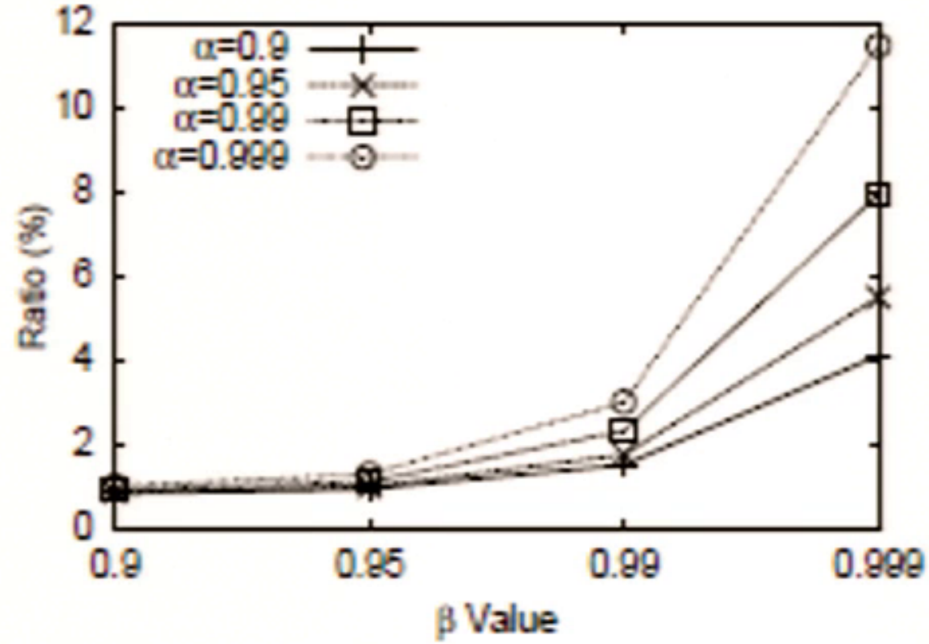


(b) NASA-HTTP Dataset

Verification Time



Verification Time



Client V.S. Server

Conclusion

- Design a probabilistic integrity verification method for outsourced privacy-preserving frequent itemset mining
- Design efficient method to construct verification objects for data perturbation based privacy preservation methods.
- Quantify the integrity guarantee probability.
- Conduct experiments to evaluate robustness and efficiency.

Thank You!

Questions?