ENTREPÔTS, REPRÉSENTATION & INGÉNIERIE des CONNAISSANCES

EDA

Laboratoire ERIC

Varunya ATTASENA
Nouria HARBI
Jérôme DARMONT

Sharing-based Privacy and Availability of Cloud Data Warehouses

Lyon 1 — UNIVERSITÉ LUMIÈRE LYON 2 — UNIVERSITÉ DE LYON — INSTITUT DES SCIENCES DE L'HOMME

Introduction

Business intelligence (BI) and **data analytics** have been an ever-growing trend in

DATA WAREHOUSE

Business

Finance Telecoms Insurance Logistics . . .

Non-business

Agriculture Medicine Health Environment . . .

Introduction

CLOUD COMPUTING

- On-Demand Self-Service
- Broad Network Service
- Resource pooling
- Rapid Elasticity
- Measured Service

Characteristic of Cloud Computing

Introduction

CLOUD COMPUTING

Move DW to Cloud Computing

DATA WAREHOUSE

- \+ High security
- \+ High performance of data analysis
- \- High costs of implements, maintenances…

Introduction

CLOUD DATA WAREHOUSE

- ✓ + **Low security**
- ✓ ? **Performance of data analysis ?**
- \+ **Low costs of implements, maintenances…**

Outline

Introduction

- Data warehouse
- Cloud computing

Problems

- Cloud data warehouse

Scheme I

- A new (m, n, t) multi secret sharing

Scheme II

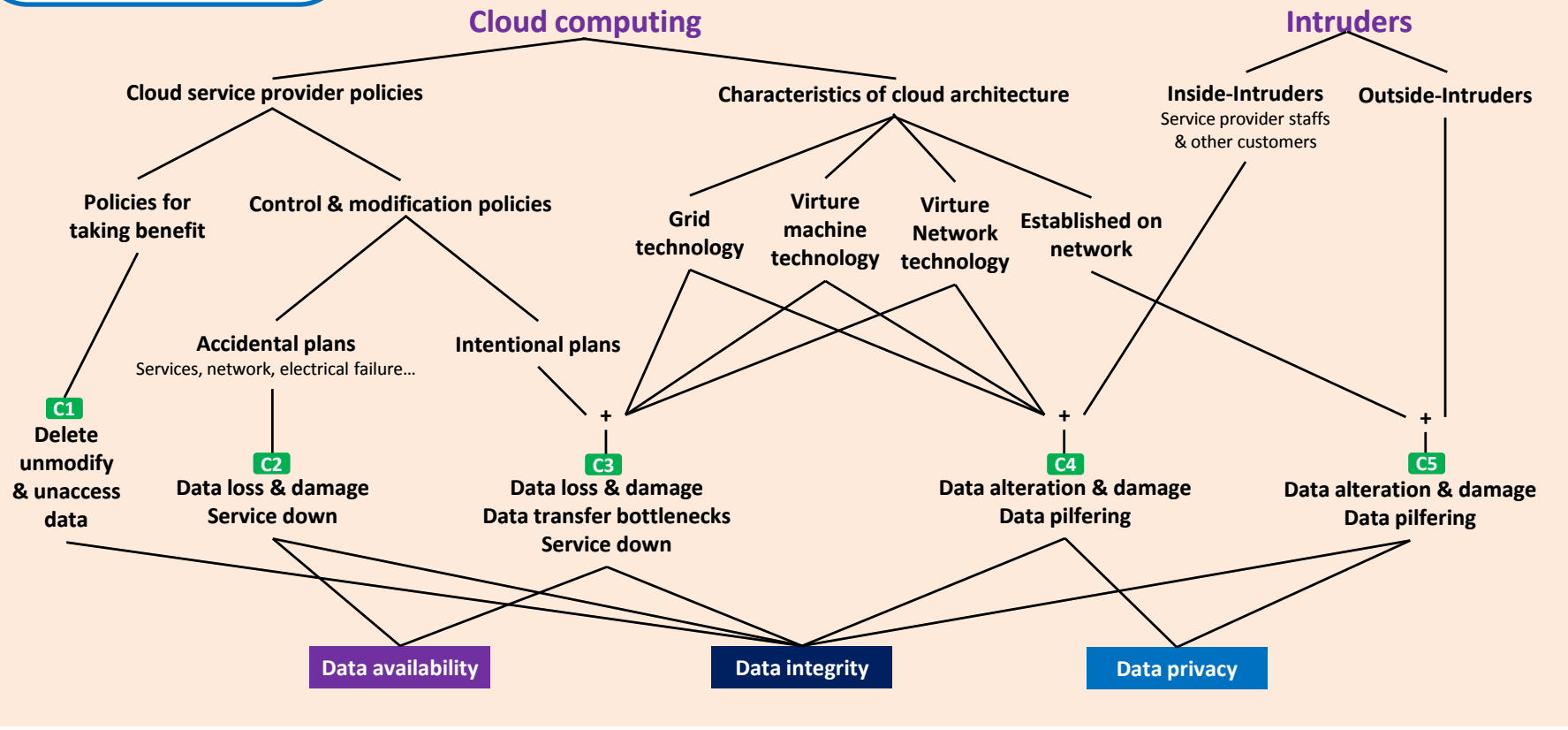
- Sharing a data warehouse in the cloud

Security analysis & performance evaluation

Conclusions

- Conclusions
- Future researches

Problems: Cloud security issues



Problems

CLOUD DATA WAREHOUSE

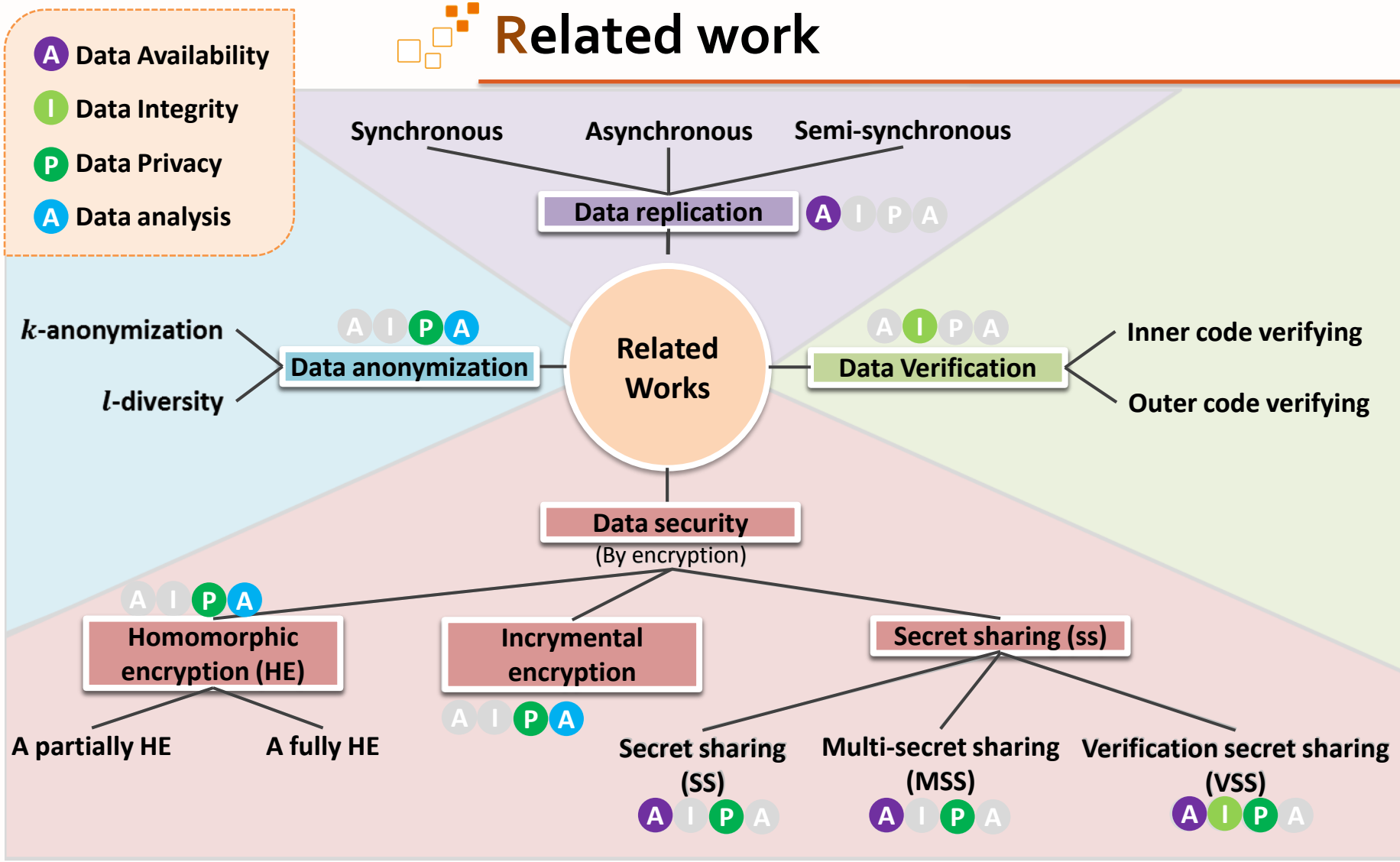
1 **Data Security**

- P Data Privacy
- A Data Availability
- I Data Integrity

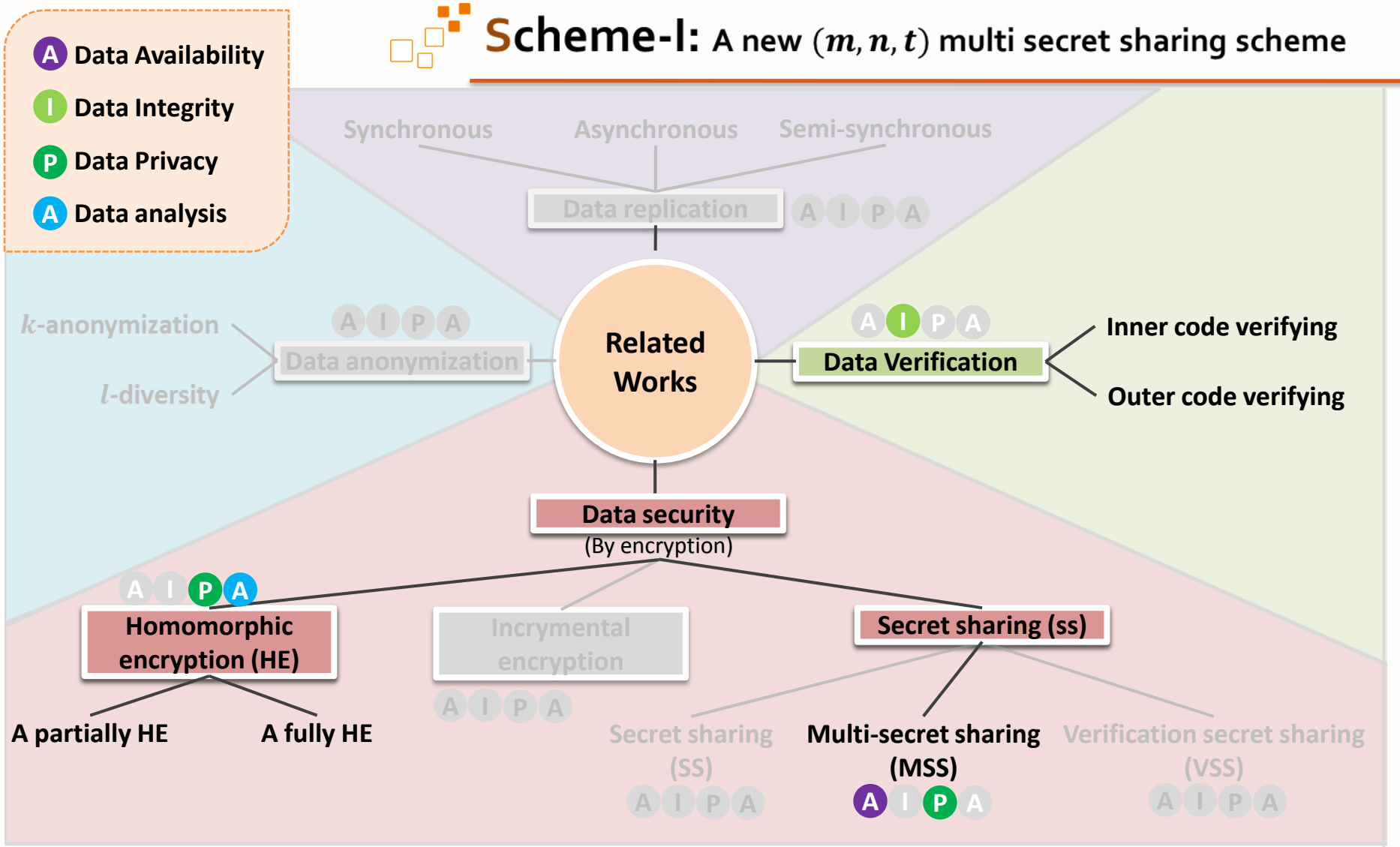
2 **Data Analysis**

- A Data Analysis

Related work



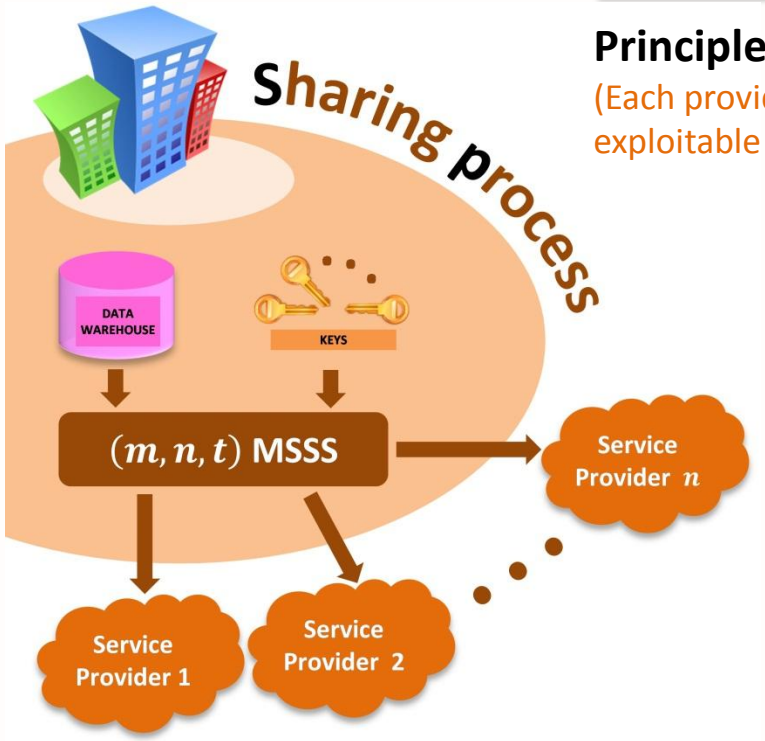
Scheme-I: A new (m, n, t) multi secret sharing scheme



Scheme-I: A new (m, n, t) multi secret sharing scheme

Principle: share data over several cloud service providers

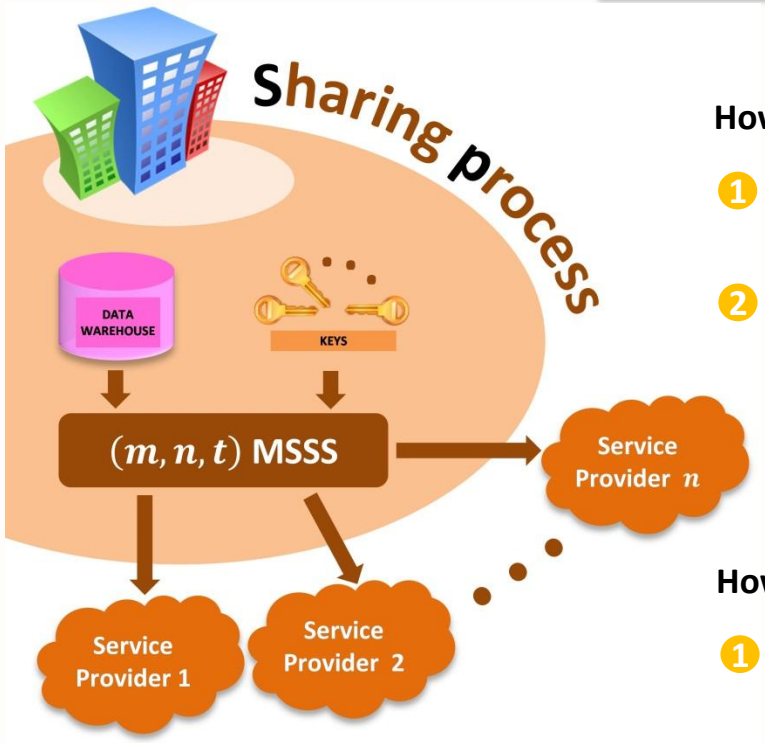
(Each provider will only store part of the data which will also not be exploitable neither by the provider nor any intruder.)



m = a number of Data.
n = a number of cloud service providers (CSPs).
t = a sufficient number of CSPs for reconstructing data.

P Data Privacy **A Data Availability** **I Data Integrity** **A Data analysis**

Scheme-I: A new (m, n, t) multi secret sharing scheme

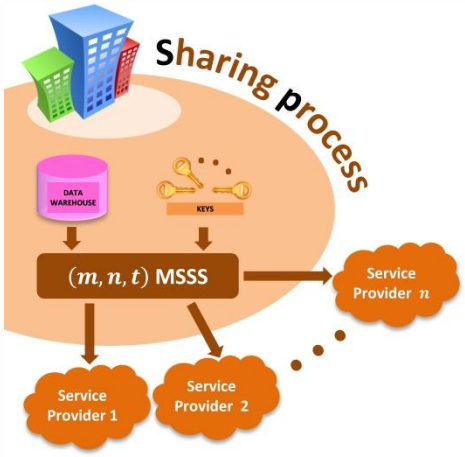


How to encrypt data?

1. data are organized into blocks.
   Each block is encrypted and decrypted all at once.
2. All data in the block are encrypted by mapping them and their signature to coefficients of a polynomial equation of degree $t - 1$.

How to verify the correctness of data?

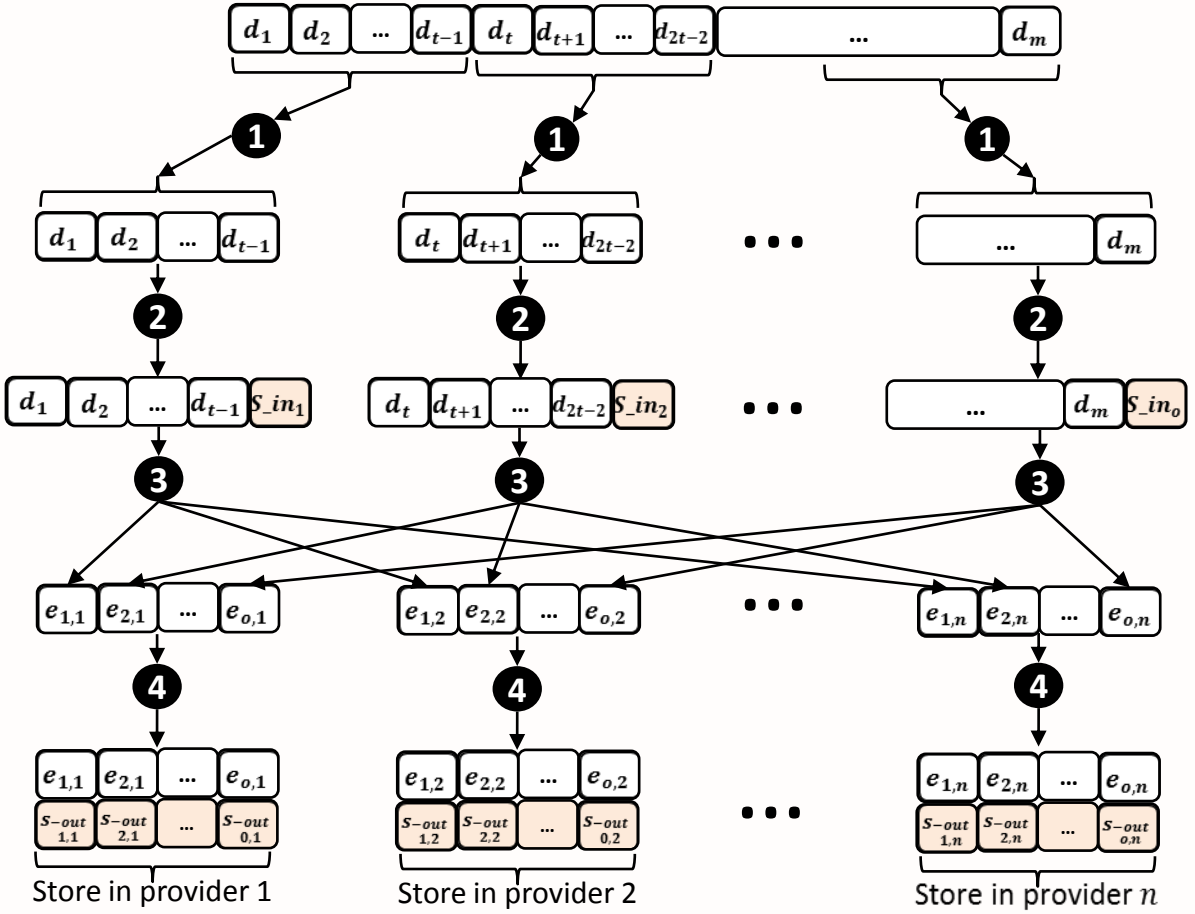
1. Create a signature s_in_j by a hash function
   (verify the trusted of providers)
2. Create a signature $s\_out_{j,k}$ by a hash function
   (reduce a cost of data transfer in reconstructing process because no error encrypted data is transferred)



Scheme-I: A new (m, n, t) multi secret sharing scheme

Sharing Process

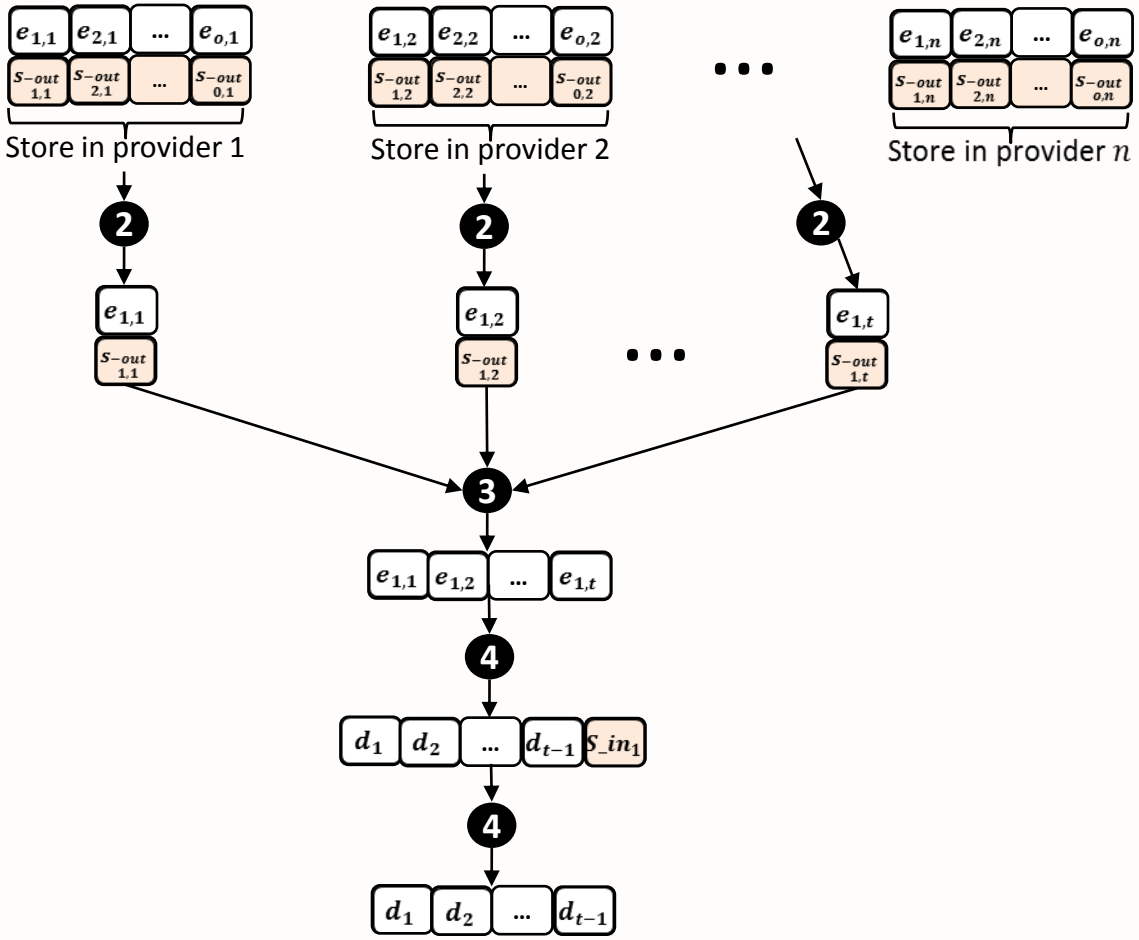
1. Data are organized into blocks.
2. Create a signature in each block.
3. Encrypt data and a signature in each block by Polynomial equation.
4. Create a signature of each encrypted data.



Scheme-I: A new (m, n, t) multi secret sharing scheme

Reconstructing Process

1. Select t CSPs from n CSPs
2. Verify a correctness of encrypted data in each CSP.
3. Transfer encrypted data to user.
4. Compute original data and a signature.
5. Verify the correctness of data.



Scheme-II: Sharing a data warehouse in the cloud

Original data

id	name	salary	sex
124	Bob	75€	M
125	Anna	80€	F

Encrypted data at CSP_1

id	name	salary	sex
124	(0,0),(10,3),(11,4)	(3,3)	(9,2)
125	(6,6),(10,3),(10,3),(0,0)	(0,0)	(10,3)

Encrypted data at CSP_2

id	name	salary	sex
124	(6,6),(6,6),(2,2)	(5,5)	(9,2)
125	(2,2),(5,5),(5,5),(0,0)	(3,3)	(7,0)

Encrypted data at CSP_3

id	name	salary	sex
124	(2,2),(0,0),(0,0)	(10,3)	(1,1)
125	(12,5),(11,4),(11,4),(0,0)	(11,4)	(7,0)

Data Analysis over shares

Can analyze data (search and aggregation operations) over shares while not decrypting all data first.

Original data			
id	name	salary	sex
124	Bob	75€	M
125	Anna	80€	F

Encrypted data at CSP_1			
id	name	salary	sex
124	(0,0),(10,3),(11,4)	(3,3)	(9,2)
125	(6,6),(10,3),(10,3),(0,0)	(0,0)	(10,3)

Select name from customer where sex='M'.
At CSP_1: Select name from customer where sex='9'.

Select avg(salary) from customer.
At CSP_1: Select avg(salary) from customer.

Security analysis and performance evaluation

Cost analysis

- **Time complexity**
- **Stored data volume**

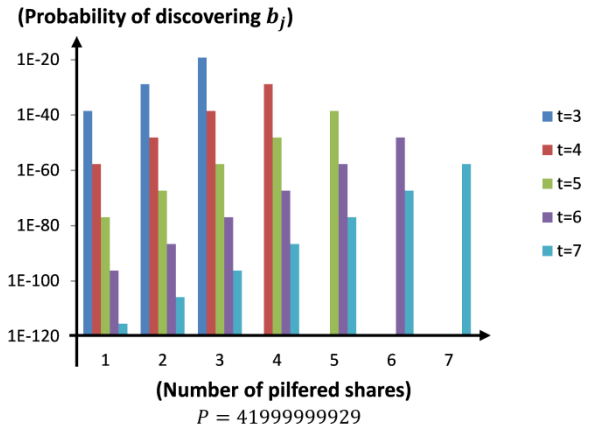
Reliability analysis

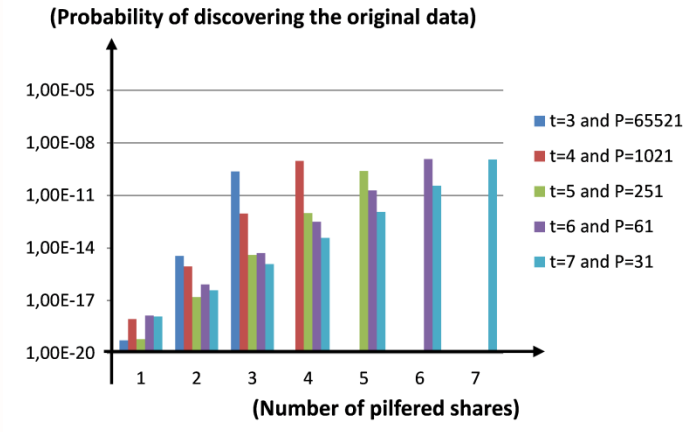
Security analysis

Security analysis

- Neither the CSP nor any intruder can decode the original data from only one share.
- It is very difficult to retrieve shares from all CSPs' by attacking them simultaneously.
- In the case that an intruder can steal shares from x CSPs such that $x \leq t$, the probability of discovering b_j is very low.

Scheme I: $\frac{1}{P^{2t-x-1}}$	**Scheme II:** $\frac{1}{p^{2t-x-1}}$	x is a number of pilfered shares.

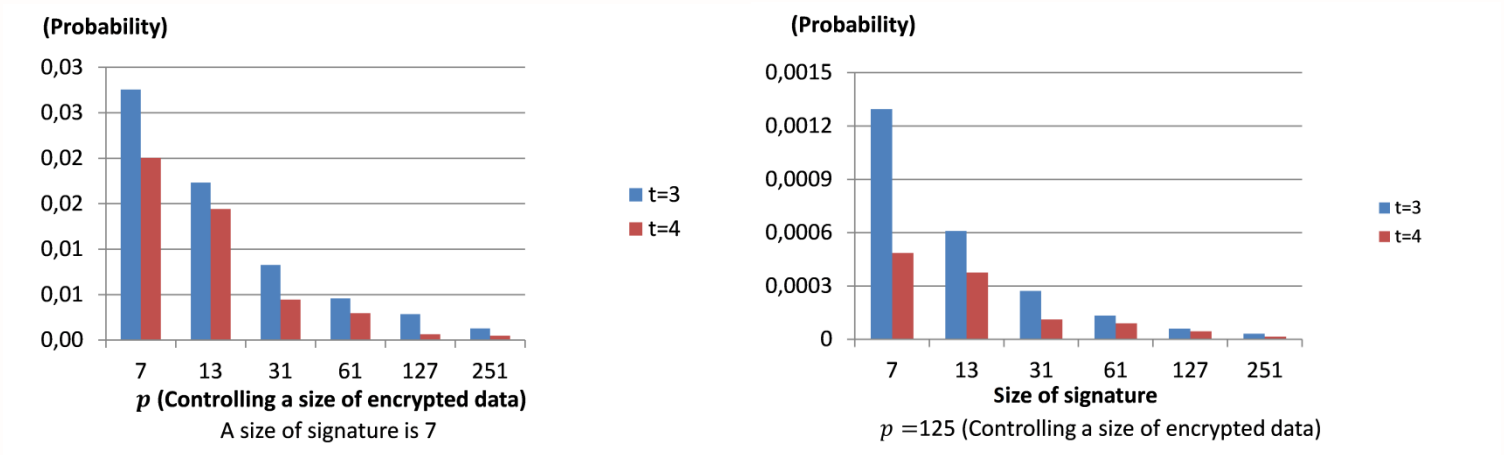




Probability of discovering an original data block from some or all shares

Reliability analysis

- **Data availability:** Our schemes guarantee the user can reconstruct *D* if *t* or more CSPs are honest and their shares are accessible.
- **Data integrity:** Our schemes can verify both the honesty of CSPs and the correctness of CSPs' shares.
- **Data recovery:** If some shares are erroneous (lost, damaged, alternative...), they are reconstructed from t other shares.



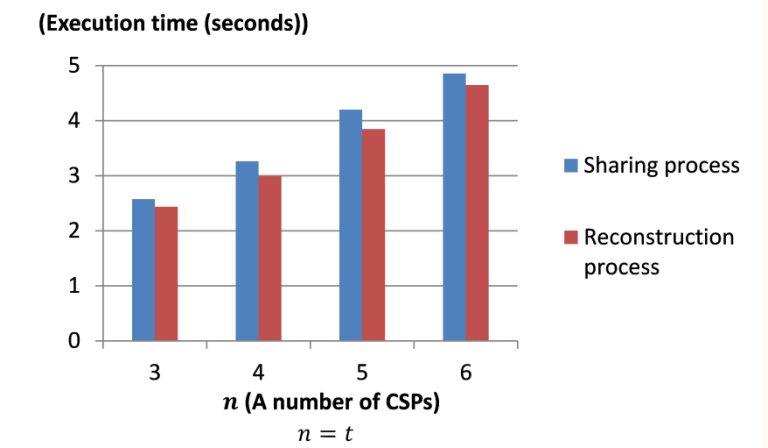
Probability of incorrect data not being detected (false negative)

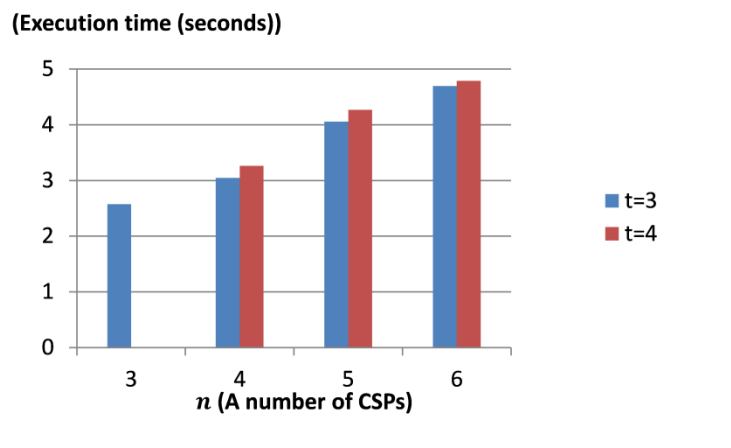
Cost analysis: Time complexity

The time complexity in both schemes

- The time complexity of the data sharing process is $O(otn)$
- The time complexity of the data reconstruction process is $O(ot^2)$

The execution time of Scheme-II: in the data reconstruction process, the execution time is about 3:04 seconds, and throughput is 336 MB per second when n = 4 and t = 3.





Probability of incorrect data not being detected (false negative)

Cost analysis: Stored data volume

The Stored data volume

- The Stored data volume in Scheme I is indeed lower than $on\|P\|$
- The Stored data volume in Scheme II is indeed lower than $on\|p\|$

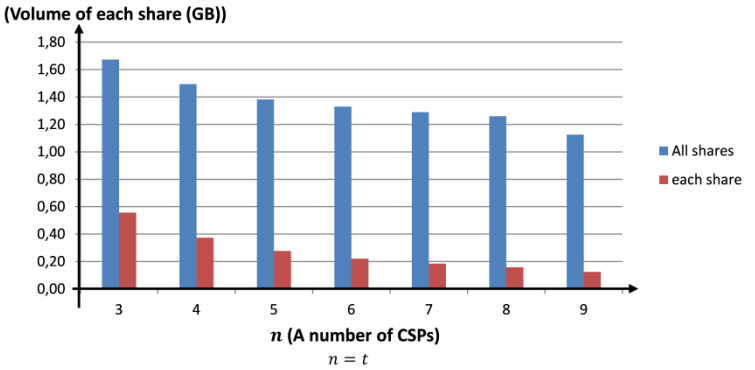
For example, with Scheme-II: 32 bits unsigned integers

(It are shared among 6 CSPs and 5 CSPs are sufficient for reconstruct them. Let $\|p\|$ = 9 bits.)

- The volume of all shares is lower than 1x6x9 = 54 bits.
- The volume of each share is lower than 1x9 = 9 bits.

By implementation of Scheme-II:

- The volume of all shares is
  greater than the volume of D but less than Dx2.
- The volume of each share is
  lower than the volume of D.



Volume of shares with Scheme-II

Conclusion

Our schemes

- A Data Availability
- I Data Integrity
- P Data Privacy
- A Data Analysis

\+

COSTS
- Time complexity
- Stored data volume
- Transferred data volume

=

Scheme-I
A new (m,n,t) multi secret sharing scheme

Scheme-II
Sharing a data warehouse in the cloud

Future researches

- A Data Availability
- I Data Integrity
- P Data Privacy
- A Data Analysis

Advanced Data Analysis
- Sophisticated aggregations
- Complex predicates (conjunction, negation, disjunction...)

\+

COST
losing data volume of all shares

Thank you