Querying Cloud Data with Confidentiality and Integrity Assurance

Jianliang Xu

Database Research Group
Hong Kong Baptist University (HKBU)
http://www.comp.hkbu.edu.hk/~db
Group Profile

- **People**
  - 3+1 faculty members
  - 9 PhD students and research/project assistants

- **Focused Research Areas**
  - Data management on new hardware
  - Data security and privacy
  - Graph and social data management
  - Mobile and spatial databases

- **Funding**
  - RGC, ITF, NSFC
  - HK$10+M grants secured in 2012-2016
Outsourcing in the Cloud

- Scalability
- Elasticity
- Self-manageability
- Pay-per-use pricing
Security Challenge #1: Integrity

- Scalability
- Elasticity
- Self-manageability
- Pay-per-use pricing

Incorrect results
- Hacking attack
- Incomplete search
- Program bug
- In favor of sponsor

Data & Algorithm

Data Owner

Cloud Service Provider (SP)

“Yellow Duck”

Client
Security Challenge #2: Confidentiality

- **Data Privacy**
  - Private asset of the data owner
  - Containing commercial intelligence information or sensitive personal information
  - Protected against the cloud server and/or the query client

- **Query Privacy**
  - 2006: AOL search data leak
  - 2013: 4 spatio-temporal locations could identify people!

**[Nature SRep, 2013]**
Roadmap

- **Motivation and Challenges**

- **Secure Data Search**
  - Private Search on Key–Value Stores [ICDE14]
  - Structure–Preserving Subgraph Queries [ICDE15]

- **Query Integrity Assurance**
  - Privacy–Preserving Query Authentication [SIGMOD12, PVLDB14]
  - Multi–Source Query Authentication [SIGMOD15]

- **Summary**
Private Search on Key–Value Store

- Example: eHR (Electronic Health Record)
  - A doctor accesses an eHR database for the test result of a patient through a patient id

- Security goal: mutual privacy
  - The doctor should receive ONLY this patient’s result, but no any other’s result
  - The SP should NOT know which test result is accessed, as it was protected by the patient–doctor confidentiality

(Name, Biotest result)
(Can Tai Man, TTF–F–FT )
(Li Chi Wai, 0.2FFS–TS)
...
Problem Definition

- A mutual privacy model:
  - Search key or returned value is **NOT** learned by the server
  - **ONLY** the value that matches the search key is returned

![Diagram showing a key-value store with search key and returned value connections.](image-url)
Preliminary 1: Homomorphic Encryption

- Plaintext space $M$ and Cipher text space $C$
  - $m_1, m_2 \in M$ and their ciphertexts $c_1, c_2 \in C$
- It holds that
  $$E^{-1}(c_1 \odot c_2) = m_1 \oplus m_2$$
- Examples: Paillier, Goldwasser–Micali (GM), RSA, El Gamal

Credit: Craig Gentry, inventor of the first fully homomorphic encryption
GT–COT: Conditional Oblivious Transfer for “Greater Than”
- The value to receiver is determined by the result of “greater–than” predicate on private inputs $x, y$ from both parties
  - The client gets $s_0$ if $x < y$, or gets $s_1$ if $x > y$
  - The predicate result is only known to the client
  - GT–COT can be implemented based on Paillier encryption

[Blake and Kolesnikov, 2004]
Applying GT–COT to Key–Search Problem

- A basic approach
  - Sort keys in ascending order
  - Invoke GT–COT for each key
  - Cost: $O(N)$ calls

- Improved version
  - Apply binary search
  - Only need $O(\log N)$ calls to GT–COT

This doesn’t work as the server needs to know the next $y$ to compare with $x$
Our Proposal: Oblivious Index Traversal

- Reducing complexity from $O(N)$ to $O(\log N)$
- Hide the $y$ value from the server by another layer of homomorphic encryption
Performance Results

(a) CPU Time

(b) Communication Bandwidth
Structure-Preserving Subgraph Queries

- **Subgraph queries** in social media search, bioinformatics, web topology...

- **Challenge**: evaluate subgraph queries at the SP while protecting the structure of query graph against SP

- **Methodology**
  - Encode query graph with cyclic group-based (CBG) encryption
  - Reduce and verify candidate mappings in encrypted domain

[IEEE ICDE 2015]
Roadmap

¬ Motivation and Challenges
¬ Secure Data Search
  ◦ Private Search on Key–Value Stores [ICDE14]
  ◦ Structure–Preserving Subgraph Queries [ICDE15]
¬ Query Integrity Assurance
  ◦ Privacy–Preserving Query Authentication [SIGMOD12, PVLDB14]
  ◦ Multi–Source Query Authentication [SIGMOD15]
¬ Summary
Authenticated Query Processing

- Empower the query client to verify the integrity of query results

- What to verify?
  - **Soundness**: all results are not tampered with
  - **Completeness**: no missing results
Query Flow vs Authentication Flow

Raw Data
- Top-K
- Range
- Skyline

Query Result

Raw Data

Crypto, Indexing

Authenticated Data Structure

Prune, Optimization

VO: Proof of Query Result
**Merkle Hash Tree**

### Merkle Hash Tree (MHT)

- **Root Node:** $N_{root}: h(N_{12}||N_{34})$
- **Internal Nodes:**
  - $N_{12}: h(N_1||N_2)$
  - $N_{34}: h(N_3||N_4)$
  - $N_1: h(d_1)$
  - $N_2: h(d_2)$
  - $N_3: h(d_3)$
  - $N_4: h(d_4)$
- **Dataset:**
  - $d_1 = 8$
  - $d_2 = 12$
  - $d_3 = 17$
  - $d_4 = 25$
- **Query:** $Q = [1, 10]$

### Service Provider

- **Database:** MHT, $\text{sig}(N_{root})$
- **Response:** $R: \{d_1, 8\}$
- **View Output:** $\{\{d_2, 12\}, N_{34}, \text{sig}(N_{root})\}$

### Data Owner

- **Client**
- **Verify**
  - **Soundness:** $8 \in [1,10];$ root
  - **Completeness:** $12, N_{34} \notin [1,10]$
Privacy–Aware Query Authentication

**Problem**: To verify query results *without* letting query client learn anything beyond

**Motivating scenarios**
- **Range Query**: Count the number of employees in a certain salary range \([α, β]\) without knowing specific salaries
- **Top-k Query**: Find new nearby friends without learning detailed locations and profiles

[ACM SIGMOD 2012, PVLDB 2014]
Building Block for Private Comparison

- To verify $v_i \geq \alpha$ without knowing $v_i$ ($v_i$ is private value)
- Basic idea: joint computing by SP and client

$Q = [\alpha, +\infty)$

$g(v_i - \alpha)$, $\text{sig}(g(v_i - L))$

$\text{sig}(g(v_i - L)), v_i$

$g(x)$ can be calculated only when $x \geq 0$

$L$: lower bound of data domain

Verify:
1. $g(v_i - L) = g(v_i - \alpha) \otimes g(\alpha_i - L)$
2. $\text{sig}(g(v_i - L))$

SP  \quad Client

DO  \quad Client
Roadmap

- Motivation and Challenges
- Secure Data Search
  - Private Search on Key–Value Stores [ICDE14]
  - Structure-preserving Subgraph Queries [ICDE15]
- Query Integrity Assurance
  - Privacy–Preserving Query Authentication [SIGMOD12, PVLDB14]
  - Multi–Source Query Authentication [SIGMOD15]
- Summary
Example: Metasearch Engines

- Combining data from multiple sources
- Providing users with a unified query interface
Verify far-away non-result values without using the whole dataset
- \( v_{12} \) proves \( v_1, v_2 \)
- \( v_{56} \) proves \( v_5, v_6 \)

Issue: \( v_1, ..., v_6 \) have sigs; \( v_{12}, v_{56} \) don’t

Challenge: aggregatable signature needed
Homomorphic Secret Sharing Seal

Content:
1. $h(\cdot)$
2. $d_{ig_i} + s_{si}$
3. seal($v$) = $G(v, s_{si}) = \left( g^{ss_i} \cdot g^{h(v(3))|h(v(2))|h(v(1))} \right)^a \mod n$

Seal design
1. Seals are “additively” homomorphic
2. Seals can be folded by the integration server
   - $v_1 = 000 \Rightarrow S_1 = G(000, s_{si})$ from $o_1$
   - $v_2 = 001 \Rightarrow S_2 = G(001, s_{si})$ from $o_2$
   - $seals(prefix(v_1, v_2)) = G(00, s_{si} + s_{si}) = S_1 \otimes S_2$
Authenticated Prefix Tree

\[ S_{root} = S_{12} \otimes S_{34} \]

\[ S_{12} = \text{seal}_1 \otimes \text{seal}_2 \]

\[ S_{34} = \text{seal}_3 \otimes \text{seal}_4 \]

Data

\[ \begin{array}{cccc}
  d_1 & d_2 & d_3 & d_4 \\
  00 & 01 & 10 & 11 \\
\end{array} \]

Q = [1, 1]

\[ Q = [1, 1] \]

Soundness: \( d_2 \in [1, 1]; S_1, S_2, S_{34} \)

Completeness: \( d_1, 1 \notin [1, 1]; \) secret
Summary

- Protection of data access is crucial in cloud settings

- It is possible and yet meaningful to support secure query processing with confidentiality and integrity assurance
  - Cryptographic-only approach cannot scale well
  - Integration with database techniques is crucial
Querying Spatial Data

- Why–Not Questions [ICDE15, ICDE16a]
  - Some object(s) unexpectedly missing from query results
  - How to **minimally** modify the initial query to revive missing object(s)?

Research Session 6A, Wednesday

- Geo–Social Group Queries [TKDE15, ICDE16b]
  - Group–based activity planning and marketing, friend gathering…
  - Challenge: **efficient** processing while considering both spatial and social constraints

TKDE Poster Session, Tuesday
Acknowledgments

- Members of HKBU DB Group
- External Collaborators
  - Christian S. Jensen (Aalborg University)
  - Prof. Sourav S. Bowmick (Nanyang Technological University)
  - Prof. Wang-Chien Lee (Penn State University)
  - Dr. Rui Chen (Samsung Research America)
- Funding Agencies
  - Research Grants Council of Hong Kong
  - Innovative Technology Fund
  - Hong Kong Scholars Program
Thank You!
References